

Statistical Techniques to Improve Software Effort Estimation for Cost Engineers

by

Tomeka S. Williams, Ph.D.

Optimum Performance Solutions (OPS) International, LLC
Empirical Research & Education Series

February 2022

Abstract

Statistical Techniques to Improve Software Effort Estimation for Cost Engineers

by

Tomeka S. Williams, Ph.D.

Optimum Performance Solutions (OPS) International, LLC
Empirical Research & Education Series
February 2022

Abstract

Since the topic of improving data quality has not been addressed for the U.S. defense cost estimating discipline beyond changes in public policy, the goal of the study was to close this gap and provide empirical evidence that supports expanding options to improve software cost estimation data matrices for U.S. defense cost estimators. The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that were referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline, and determine which theories rendered incomplete and missing data sets in a single data matrix most reliable and complete under eight missing value percentages. A quantitative pre-experimental research design, a one group pretest-posttest no control group design, empirically tested and measured the predictive accuracy of traditional missing data theory techniques typically used in non-cost estimating disciplines. The results from the pre-experiments on a representative U.S. defense software cost estimation data matrix obtained, a nonproprietary set of historical software effort, size, and schedule numerical data used at Defense Acquisition University revealed that single and multiple imputation techniques were two viable options to improve data quality since calculations fell within 20% of the original data value 16.4% and 18.6%, respectively. This study supports positive social change by investigating how cost estimators, engineering economists, and engineering managers could improve the reliability of their estimate forecasts, provide better estimate predictions, and ultimately reduce taxpayer funds that are spent to fund defense acquisition cost overruns.

Table of Contents

List of Tables	v
List of Figures	ix
Chapter 1: Introduction to the Study.....	1
Background of the Study	5
Problem Statement	7
Purpose of the Study	9
Research Question and Hypotheses	12
Theoretical Foundation	14
Nature of the Study	16
Definitions.....	19
Assumptions.....	24
Scope and Delimitations	25
Limitations	26
Significance of the Study	28
Significance to Theory	29
Significance to Practice.....	30
Significance to Social Change	30
Summary and Transition.....	31
Chapter 2: Literature Review	35
Literature Search Strategy.....	37
Theoretical Foundation	40

Literature Review.....	45
U.S. Public Policy Requires Data Reliability/Completeness in Cost	
Estimation	46
Effect of U.S. Public Policy on Cost Estimation Data	
Reliability/Completeness	55
Gap: Expanding Options to Handle the Estimator’s Unreliable/Incomplete	
Data Problem	57
How Other Disciplines Handle Unreliable/ Incomplete Data Problems	64
Could Missing Data Theory Improve Data Reliability/Completeness?	65
Summary and Conclusions	66
Chapter 3: Research Method.....	70
Research Design and Rationale	71
Methodology	74
Population	74
Sampling and Sampling Procedures	77
Procedures for Data Collection (Purposive Sample of Archival Data)	79
Intervention (One Group Pretest-Posttest Design/Pre-experimental).....	79
Archival Data	83
Instrumentation and Operationalization of Constructs (IBM SPSS 25)	83
Intervention Studies or Those Involving Manipulation of Independent	
Variables	84
Data Analysis Plan.....	86

Threats to Validity	90
External Validity	91
Internal Validity	91
Construct Validity	92
Ethical Procedures	93
Summary	94
Chapter 4: Results	96
Data Collection	99
Treatment and Intervention Fidelity	105
Study Results	115
First Evaluation Measure to Determine Predictive Accuracy	115
Second Evaluation Measure to Test Main Effects & Interactions	119
Summary	126
Chapter 5: Discussion, Conclusions, and Recommendations	129
Interpretation of Findings	130
Limitations of the Study	131
Recommendations	133
Implications	134
Significance to Theory	135
Significance to Practice	136
Significance to Social Change	137
Conclusions	137

References	141
Appendix A: Closest Predictive Accuracy Results Per Data Set.....	162
Appendix B: Two-Way Repeated Measures ANOVA in SPSS Selection	207
Appendix C: Select SPSS Outputs from Two-Way Repeated Measures ANOVA	213

List of Tables

Table 1. Select U.S. Public Policy Criteria and Requirements to Produce Reliable and Complete Cost Estimates	48
Table 2. Summary from Data Used within a Synthetic and Complete U.S. Representative Software Cost Estimation Data Matrix Used at Defense Acquisition University (DAU)	59
Table 3. DoD Major Automated Information Systems (MAIS) List.....	75
Table 4. Selected Analogous Programs from DAU Data Matrix Based on Application Type (e.g., Operational Environment, Development Paradigm and Phase)	100
Table 5. Selected Numerical Data Sets from DAU Data Matrix	103
Table 6. Summary of Closest Missing Data Theory (MDT) Technique Predictive Accuracy Results for Twenty-Eight Empirically Tested U.S. Defense Software Cost Estimating Data Types	116
Table 7. Degree to Which Missing Data Theory Techniques Can Solve the U.S. Cost Estimators' Unreliable and Incomplete Data Problem Based on Approximation Error	119
Table AI. Closest Missing Data Theory Predictive Accuracy in Data Set 1 (DS1) Experiments for Number of External Interfaces Types of Data.....	162
Table A2. Closest Missing Data Theory Predictive Accuracy in Data Set 2 (DS2) Experiments for Initial Software Lines of Code (SLOC)-New Types of Data...	164
Table A3. Closest Missing Data Theory Predictive Accuracy in Data Set 3 (DS3) Experiments for Initial SLOC Modified Types of Data.....	165

Table A4. Closest Missing Data Theory Predictive Accuracy in Data Set 4 (DS4)	
Experiments for Initial SLOC Reused Types of Data.....	167
Table A5. Closest Missing Data Theory Predictive Accuracy in Data Set 5 (DS5)	
Experiments for Final SLOC – New Types of Data.....	168
Table A6. Closest Missing Data Theory Predictive Accuracy in Data Set 6 (DS6)	
Experiments for Final SLOC – Modified Types of Data.....	170
Table A7. Closest Missing Data Theory Predictive Accuracy in Data Set 7 (DS7)	
Experiments for Final SLOC – Reused Types of Data.....	171
Table A8. Closest Missing Data Theory Predictive Accuracy in Data Set 8 (DS8)	
Experiments for Re-Design/ Design Modified Effort (DM) % - Modified Types of Data.....	173
Table A9. Closest Missing Data Theory Predictive Accuracy in Data Set 9 (DS9)	
Experiments for Re-Code/ Code Modified Effort (CM) % - Modified Types of Data.....	174
Table A10. Closest Missing Data Theory Predictive Accuracy in Data Set 10 (DS10)	
Experiments for Re-Test/ Integration Modified Effort (IM) % - Modified Types of Data.....	176
Table A11. Closest Missing Data Theory Predictive Accuracy in Data Set 11 (DS11)	
Experiments for Design Modified (DM) % - Reused Types of Data.....	177
Table A12. Closest Missing Data Theory Predictive Accuracy in Data Set 12 (DS12)	
Experiments for Code Modified (CM) % - Reused Types of Data.....	179
Table A13. Closest Missing Data Theory Predictive Accuracy in Data Set 13 (DS13)	

Experiments for Integration Effort (IM) % - Reused Types of Data.....	180
Table A14. Closest Missing Data Theory Predictive Accuracy in Data Set 14 (DS14)	
Experiments for Final Software Requirements Analysis Effort Hours Types of Data.....	183
Table A15. Closest Missing Data Theory Predictive Accuracy in Data Set 15 (DS15)	
Experiments for Final Software Architectural Design Effort Hours Types of Data.....	184
Table A16. Closest Missing Data Theory Predictive Accuracy in Data Set 16 (DS16)	
Experiments for Final Software Detailed Design Effort Hours Types of Data.....	186
Table A17. Closest Missing Data Theory Predictive Accuracy in Data Set 17 (DS17)	
Experiments for Final Software Construction Effort Hours Types of Data.....	187
Table A18. Closest Missing Data Theory Predictive Accuracy in Data Set 18 (DS18)	
Experiments for Final Software Integration Effort Hours Types of Data.....	189
Table A19. Closest Missing Data Theory Predictive Accuracy in Data Set 19 (DS19)	
Experiments for Final Software Qualification Testing Effort Hours Types of Data.....	190
Table A20. Closest Missing Data Theory Predictive Accuracy in Data Set 20 (DS20)	
Experiments for Final Software Documentation Management Effort Hours Types of Data.....	192
Table A21. Closest Missing Data Theory Predictive Accuracy in Data Set 21 (DS21)	

Experiments for Final Software Configuration Management Effort Hours Types of Data.....	193
Table A22. Closest Missing Data Theory Predictive Accuracy in Data Set 22 (DS22)	
Experiments for Final Software Quality Assurance Effort Hours Types of Data.....	195
Table A23. Closest Missing Data Theory Predictive Accuracy in Data Set 23 (DS23)	
Experiments for Final Software Verification Effort Hours Types of Data ,,,.....	196
Table A24. Closest Missing Data Theory Predictive Accuracy in Data Set 24 (DS24)	
Experiments for Final Software Validation Effort Hours Types of Data.....	198
Table A25. Closest Missing Data Theory Predictive Accuracy in Data Set 25 (DS25)	
Experiments for Final Software Review Effort Hours Types of Data.....	199
Table A26. Closest Missing Data Theory Predictive Accuracy in Data Set 26 (DS26)	
Experiments for Final Software Audit Effort Hours Types of Data.....	201
Table A27. Closest Missing Data Theory Predictive Accuracy in Data Set 27 (DS27)	
Experiments for Final Software Problem Resolution Effort Hours Types of Data.....	202
Table A28. Closest Missing Data Theory Predictive Accuracy in Data Set 28 (DS28)	
Experiments for Final Cybersecurity Effort Hours Types of Data.....	205

List of Figures

Figure 1. Statistical method to perform this One Group Pretest-Posttest Design.....	12
Figure 2. Statistical method to perform this One Group Pretest-Posttest Design.....	86
Figure 3. Statistical method to perform this One Group Pretest-Posttest Design.....	107
Figure 4. Removed-at-Random-Data-Value Positions to Create the Artificially Induced Missing Data Problem at Eight Percentage Levels of Missingness.....	112
Figure 5. Systematic Approach to Artificially Induce the Missing Data Problem at Eight Levels of Missingness for Three Missing Data Theory (MDT) Technique Treatments.....	114
Figure 6. Plot of the Results to Assess Interaction Between Approximation Error (Dependent Variables) of the Actual/Pretest (1) and Computed Posttest (2) Value Means	122
Figure 7. Interaction Analysis for the Two-Way Repeated Measures ANOVA	123
Figure 8. Statistical Significance and Main Effect from the Two-Way Repeated Measures ANOVA	125
Figure B1. Select Analyze, General Linear Model, and Repeated Measures Screen.....	207
Figure B2. Select Within-Subjects Variables and Between-Subjects Factors.....	208
Figure B3. Define Profile Plots to Determine if the Means are Equal on each Missing Data Theory Technique.....	209
Figure B4. Define Post Hoc Tests for the Independent Variables.....	210
Figure B5. Define Estimated Marginal Means.....	211
Figure B6. Define Options to Analyze.....	212

Figure C1. Within-Subjects Factors Coded in SPSS.....	213
Figure C2. Between-Subjects Factors Coded in SPSS.....	213
Figure C3. Box’s Test of Equality of Covariance Matrices.....	215
Figure C4. Mauchly’s Test of Sphericity.....	216
Figure C5. Levene’s Test of Equality of Error Variance.....	216

Chapter 1: Introduction to the Study

From the perspective of U.S. public policy statutes and regulations, the U.S. defense cost estimating discipline, and the current *Business—Cost Estimating* curriculum at Defense Acquisition University (DAU), there is a lack of instruction in which cost estimators, engineering economists, and engineering managers can apply to handle the unreliable and incomplete engineering project *data matrix* problem they face (DAU, 2018a; GAO, 2009, 2020; International Cost Estimation and Analysis Association [ICEAA], 2019). According to a U.S. defense based *Joint Agency Cost Estimating Relationship (CER) Handbook* (2018), “data sets with missing and incomplete data” is a data analysis challenge and states that the “best course of action is to first attempt to remedy the problem by collecting more data, finding the information from the collected data set, and determining the cause of the unusual observations, respectively” (p. 221). This government document also acknowledges that it is “not always possible to correct such errors” and that it is important for estimators to understand the implications of these challenges, and to proceed with their analysis under caution (Joint CER Handbook, 2018, p. 221). The literature does not inform how cost estimators who leverage the defense *Business—Cost Estimating* curriculum at DAU directly handles the unreliable and incomplete engineering project data matrix problem other than through recognizing the problem through defense government documents and making changes to public policy (GAO, 2020). Because the topic of improving data quality as it relates to data incompleteness has never been addressed for U.S. defense cost estimators and the cost estimating discipline beyond describing the problem or making changes in its public

policies, research was needed to fill the gap in the literature to investigate if the use of *hands-on-treatment-options* could improve software cost estimation of data matrices for this population in the society (10 U.S. Code § 1746, 2012; 10 U.S. Code § 2334, 2017; Morin, 2017). Hands-on-treatment-options could provide the ability to use missing data theory techniques to teach cost estimators ways in which they could directly handle unreliable and incomplete data within the cost estimation discipline. This includes but is not limited to applying missing data theory techniques such as complete case analysis (listwise delete), direct imputation (single or multiple), model-based imputation (full information maximum likelihood), and machine learning methods (García-Laencina et al., 2010).

My research specifically honed-in on the area of software cost estimation because it was the cost estimation topic most commonly found in scholarly peer-reviewed journal articles, conference proceedings, and academic books in respect to this discipline (see Boehm, 1981; Idri et al., 2016b; Jing et al., 2016; Jones, 2007; Strike et al., 2001). Software cost estimation is the process taken to quantify the cost of expected labor effort, lines of code, and calendar time required to develop a software engineering project (Wani et al., 2019).

This gap-mitigating research was needed and can be used to inform ways in which U.S. defense cost estimators could have empirical evidence that supports expanding options for them as individuals to single-handedly address the unreliable and incomplete data problem beyond the sole dependence of public policy changes as referenced in Morin's (2017) and the Department of Defense (DoD) cost analysis data

improvement plan. These hands-on-treatment-options will model and simulate the conditions which cost estimators face when they are sitting in front of their computers attempting to create an estimate with imperfect data. Beyond public policy changes, cost estimators could have additional hands-on-treatment-options to use missing data theory techniques within this discipline for the first time. This original research could change how estimators are taught to conduct software cost estimation activities when the unreliable and incomplete data problems create imperfect *data sets* to use, a real-world data quality issue (see Morin, 2017). A quantitative research design, specifically a one group pretest-posttest no control group/pre-experimental design, measured the level of predictive accuracy of traditional missing data theory techniques treatments applied to 28 data sets from a single group data matrix (Thyer, 2012; see Campbell & Stanley, 1963; Cook & Campbell, 1979; Reichardt, 2019; Shadish et al., 2002). *Predictive accuracy* is the operational term used to describe how close the error approximation is between the ground truth data sets, the a priori value, as compared to its posteriori value after applying missing data theory technique treatments (Little & Rubin, 2020; Twala et al., 2006).

This study provided an opportunity for societal change by investigating how cost estimators, engineering economists, and engineering managers could benefit from additional options that directly improve data incompleteness, create better estimate predictions, and ultimately reduce taxpayer funds that are spent on defense acquisition cost overruns (Schwartz & O'Connor, 2016). Missing data theory techniques have been applied and used by many professionals in other disciplines since missing data theory

was introduced in 1987 (Little & Rubin, 1987). It continues to maintain its relevance to improve data quality by making data sets complete in a multitude of disciplines (Little & Rubin, 2020). The results of this research's one group pretest-posttest no control group/pre-experimental design treatment results report out to what degree does missing data theory techniques accurately impute data values compared to their original true and complete values. The level of predictive accuracy measured by my calculations displayed the delta between the pretest and posttest values, focusing on approximation error expressed as both a number and a percentage. The difference between the pretest and posttest numerical values informs other researchers of the Business—Cost Estimating discipline and helps them better understand to what degree could missing data theory render data sets complete. The dependent variables of absolute error and relative error are the two measures of predictive accuracy used in this study. There were three independent variables used: the different percentage levels of missingness created, the categorical name of data set type chosen, and the missing data theory techniques chosen and applied to a synthetic, representative U.S. defense cost estimation matrix for which all data values were initially completely in place and known (i.e., a nonproprietary set of software effort and size estimation complete numerical data). In addition to measuring the level of predictive accuracy, I measured the main effects and interactions between the independent variables to test for significance by conducting ANOVA testing. The pre-experimental findings and results demonstrated that missing data theory techniques could be a viable option to correct imperfect data that is unreliable or incomplete with a data value that is closer to the ground truth of the original numerical values. The purpose of

this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages.

Chapter 1 includes the problem and purpose statement of this empirical study and addresses the gap in the DoD cost estimation discipline literature to support future improvements in both the engineering economics and management fields of study. This chapter also contains the specific research goal, objectives, and the scope of this research. Moreover, this chapter describes the motivation of this research project to improve the state of practice and bring about social change (Govinfo, 2020).

Background of the Study

When the GAO (1972) studied the problem that cost estimates were not reliable, they commented that, "historical cost data used for computing estimates were sometimes invalid, unreliable, or unrepresentative" (p. 1). Thirty-seven years later, GAO (2009) stated the same problem and attempted to provide additional guidance and structure for cost estimators to use more reliable data matrices as a fix. Unfortunately, the guidance from this government document was not comprehensive and did not address what a U.S. cost estimator could do, in a hands-on manner, to handle data matrices that are unreliable or incomplete. Furthermore, it did not address what options cost estimators have available

to them to handle the unreliability and incompleteness of their data matrices for different types of engineering-based acquisition projects and programs.

Forty-eight years since 1972, GAO (2020) published an update to its 2009 government document titled, *GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs* (Report No. GAO-09-3SP) and acknowledged that “developing reliable cost estimates is crucial for realistic program planning, budgeting, and management” (p. 3). This government document was developed to close the gap in the field by documenting “generally accepted best practices for ensuring reliable cost estimates (applicable across government and industry)” and represents what has been done at the U.S. government level in respect to “processes, procedures, and practices” that have been used in the defense cost estimation body of knowledge (GAO, 2020, p. 3). This government document supports the claim that improvements are still desired, a gap needs to be closed, and there remains a lack of scholarly research in the Business—Cost Estimating discipline that addresses what additional options U.S. defense cost estimators must handle data matrices that may be unreliable or incomplete (GAO, 2020).

In the government publication for DoD cost analysis data improvement, Morin (2017) stated that several cost estimating oversight organizations that collect and store software cost estimating data, such as the Office of the Secretary of Defense for Cost Analysis and Program Evaluation (OSD CAPE), could benefit by improving the data quality problem by “closing data gaps” (p.1). In the literature, others who leverage and assess software effort data in non-U.S. defense sectors agree that effort estimation is an

important step in software projects, acknowledging that missing data occurs in real world data collection, and have found imputation strategies to be helpful to improve their software effort estimation performance (Jing et al., 2016; Qi et al., 2017). Unfortunately, it is common for many U.S. defense data matrices to not have complete data sets to support the development of credible cost estimates, thus the drive by Morin (2017), the former OSD CAPE director, to “provide cost, acquisition, and resource allocation organizations with data required for better analysis and decision-making” (p. 1). All too often, even if organizations obtain all project data, the data are typically incomplete (Jing et al., 2016). Within the context of the defense cost estimating body of knowledge, and current U.S. federal curriculum at DAU, how to handle an incomplete physical project data matrix has never been addressed and is needed to support the public policy requirement for reliable and complete cost data to produce credible cost estimates (DAU, 2018a; GAO, 1972, 2009, 2020; ICEAA, 2019; Morin, 2017).

Problem Statement

There are over 50 federal public policies, statutes, and regulations in place today that apply to the Business—Cost Estimating discipline that is required to produce reliable cost estimates (DAU, 2018b; GAO, 2009, 2020). The general management problem is that despite this, cost estimators do not always have reliable and complete data sets to use when they attempt to forecast life-cycle costs for a myriad of engineering-based acquisition projects and programs and may sometimes forecast costs inaccurately that engineering managers depend on (GAO, 2009, 2020; Jorgensen, 2006; Morin, 2017). Consequently, in 2015, cost estimate growth was reported as cost overruns within DoD’s

Major Defense Acquisition Programs at \$468 billion, up from \$295 billion in 2008 (Deloitte, 2016). Other studies declared that individual DoD engineering-based acquisition projects and programs experienced cost overruns as high as 40% and were projected to overrun closer to 51% by the year 2020 (Christensen, 1993; Dabkowski & Valerdi, 2016; Deloitte, 2016; Valerdi et al., 2015). These costs overrun statistics support the currency and relevancy that the lack of tools to handle incomplete and faulty data is a current, real-world problem so severe that a DoD cost analysis data improvement effort was started (Morin, 2017). According to Morin (2017), "reliable and comprehensive cost data is essential to produce credible cost estimates as required in both statute and regulation" (p. 1). This supports the U.S. defense cost estimator's need for cost data improvements (Morin, 2017). Multiple credible sources have noted how important it is to have reliable and comprehensive cost data for the multi-discipline of cost estimation which spans the business, engineering economics, software, and systems engineering disciplines (DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013; Jorgensen, 2006; Morin, 2017; Newnan et al., 2004; Parnell, 2017).

The specific management problem is that there is a lack of research into the techniques to handle the unreliable and incomplete data problem. Consequently, the research problem is that there is a lack of knowledge and understanding among cost estimators about what options they have to improve the data quality of data sets with limited, incomplete, or unreliable data, which prevents them from forecasting accurately the life-cycle costs for a myriad of engineering-based acquisition projects and programs (DAU, 2018a; GAO, 2009, 2020; Morin, 2017). Morin (2017) stated that research into

data improvements were needed and specifically called out the need to “improve analyst productivity”, “close data gaps”, and ultimately incorporate data quality procedures through policy or guidance to make cost analysis data more reliable and complete (p.1). In other disciplines, there are various data improvements used by researchers and analysts to handle data matrices that include unreliable, incomplete, and missing data values (Allison, 2002; Enders, 2010; Graham, 2012; Little & Rubin, 2002, 2020). Though there is literature on ways to handle missing values using missing data theory in other disciplines, there is a gap that needs to be addressed within the current research related to the U.S. defense cost estimation body of knowledge that describes how defense cost estimators could handle the unreliable and incomplete data quality problem (Brown & White, 2017; DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013; GAO, 2009, 2020; Mislick & Nussbaum, 2015).

Purpose of the Study

The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories’ results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. The three independent variables used for this study were the different percentage levels of missingness created (independent variable 1), the category title of the data set type (independent variable 2), and the traditional missing data theory techniques (independent variable 3). The two dependent

variables used for this study were the absolute errors and relative errors calculated from the pre-experimental treatments derived from the data sets' pretest and posttest numerical values. Differences revealed from the absolute error and relative error groups were assessed by ANOVA testing. I used eight different percentages for missing values (diminished completeness) with three treatments on the randomly selected subset of a purposive sample of 30 out of 50 analogous and synthetic software development programs. Each program was characterized across 28 numerical data sets. Due to the removed-at-random value selection to test and measure at eight different levels of missingness, each of the data sets had missing data theory treatments applied to fill in incomplete data 56 times, resulting in a total of 4,704 ($3 \times 56 \times 28$) pre-experimental treatments.

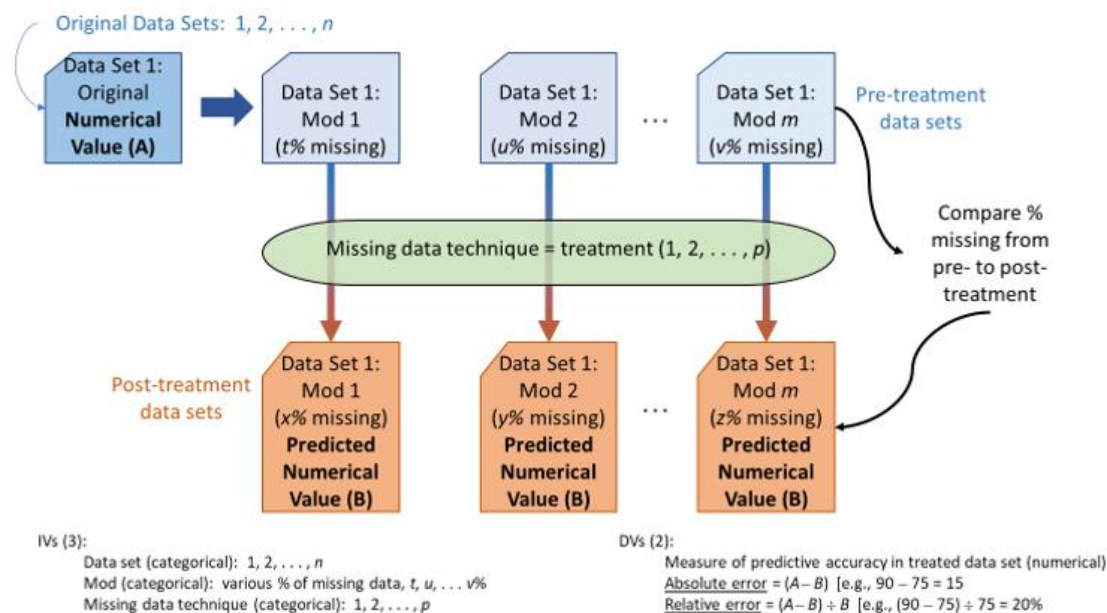
By conducting this research, I closed a gap in the U.S. cost estimation discipline and added to the research, knowledge, and understanding which serve as rationale for employing additional options for cost estimators to perform more reliable and complete cost estimation products for major DoD engineering-based acquisition projects and programs. The results of this test provide U.S. defense cost estimators with an evaluation of which additional set of options can handle the unreliable and incomplete data problem when building a cost estimate (see DAU, 2018a; GAO, 2009, 2020; Morin, 2017).

Two levels of measurement, absolute error and relative error, were used to measure the predictive accuracy of missing data approaches. Each of the 28 data set types (independent variable 2) used in this study had data values removed-at-random and at various percentages (independent variable 1) to create a simulation of the missing data

problem applied to a representative defense cost estimating data. The removed-at-random-data-values from the data matrix were operationally named the “Original Numerical Value” (pretest value) (Research Randomizer, 2020). The predicted value created because of applying the missing data theory technique (independent variable 3) were operationally named the “Predicted Numerical Value” (posttest value) for each run of the experiment. The absolute error and relative error outcome variable were the delta values calculated, the error approximation values, between the “Original Numerical Value” (pre-experiment’s pretest value) and the “Predicted Numerical Value” (pre-experiment’s posttest value) to determine each missing data theory technique’s predictive accuracy (Idri et al., 2015a, 2015b, 2016a, 2016b, 2016c; Twala et al., 2006). Figure 1 provides a graphical depiction of the study.

Figure 1

Statistical Method to Perform this One Group Pretest-Posttest Design



Research Question and Hypotheses

The research question (RQ) for this study was intended to investigate what the predictive accuracy was from various missing data theory techniques when applied to a defense cost estimating data matrix: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? The null and alternative hypotheses that used to answer the RQ were derived from the results of the sole data matrix using the one group pretest-posttest (no control group/pre-experimental) design. The calculated measure of predictive accuracy (e.g., error approximation value) provided a table of before and after average absolute and average relative error values because of the applied three treatments of

missing data theory techniques. There were 28 data set types tested, each comprised of only 30 out of 50 analogous and synthetic software development programs as the foundation for this pre-experiment. After which, analysis of variance (ANOVA) test was conducted to explain the interaction of this study's two dependent variables using the following null and alternate hypotheses:

H_01 : There are no significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, the Complete Case Analysis/ Listwise Delete approach?

H_{a1} : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, Complete Case Analysis/ Listwise Delete approach? The means are not equal.

H_02 : There are no significant differences evident between the data set's mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 2, a Single Imputation approach?

H_{a2} : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 2, a Single Imputation approach? The means are not equal.

H_{03} : There are no significant differences evident in the data sets' mean absolute error and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 3, the Multiple Imputation approach?

H_{a3} : There are significant differences evident between the data sets' mean absolute error and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 3, the Multiple Imputation approach? The means are not equal.

Theoretical Foundation

The theoretical framework that grounded this pre-experimental study was missing data theory (see Allison, 2002; Graham, 2012; Little & Rubin, 1987, 2002, 2020; Rubin, 1976). The intent of this study was to ascertain how effective the missing data techniques' measure of predictive accuracy was when applied to the defense cost estimating discipline's variant of the missing data problem, each technique based on missing data theory. Since this theory addressed data completeness, it was most appropriate to use since several years of research have validated that the theory provides effective techniques to fill data gaps in many non-U.S. defense cost estimation disciplines (see Allison, 2002; Enders, 2010; Graham, 2012; Little & Rubin, 2002, 2020).

Disciplines that have leveraged missing data theory can be found in the social sciences, health, pharmaceutical industry and practically any industry that requires an assessment

of data to inform decision-makers (see Allison, 2002; Blankers et al., 2010; Little & Rubin, 2002, 2020).

This theory was a starting point to determine which missing data theory technique (independent variable 3) could improve the state of the U.S. defense cost estimating data reliability and completeness problem and is discussed further in Chapter 2. This framework allowed me to look at missing data through a narrower lens by testing various traditional missing data theory techniques (e.g., complete case analysis/listwise delete, single imputation, and multiple imputation) to inform and update how cost estimators could handle and treat areas of missing data. The theoretical framework provided a basis for answering the RQ: To what degree can traditional missing data theory techniques accurately solve cost estimator's and engineering manager's unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? This research can inform the U.S. defense cost estimation discipline about additional options to improve data quality beyond policy incorporation and could offer these options as a new topic to include in future courses and curriculum. Furthermore, this research could inform how this topic is addressed within other academic books and journals involving engineering economics/management, software engineering economics, machine learning data preprocessing, cost analysis, financial decision sciences, data preparation (e.g., data inclusion/exclusion and data cleansing), data mining, and of course Business—Cost Estimating at DAU (see Boehm, 1981,1984, 2002; DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013; Gautam & Ravi, 2015; Nagashima & Kato, 2019; Van Hulse & Khoshgoftaar, 2014; Williams & Barber, 2011).

Nature of the Study

The nature of this study was a quantitative method approach to inquiry using a pre-experimental study design. Various experimental study designs (pre-, quasi, or true experiments) are a proven approach to comparatively test and measure the predictive accuracy of missing data theory techniques using a pretest-posttest no control group design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Crammer, 2018; Kirk, 2013; Reichardt, 2019; Shadish et al., 2002; Shek & Zhu, 2018; Singleton & Strait, 2010). To elucidate how effective each missing data theory technique was, a publicly sourced nonproprietary data matrix was obtained and manipulated to experiment on 28 out of 34 ratio scale/numerical software cost estimation data set types (independent variable 2) used within the U.S. defense cost estimating discipline from a representative data matrix. In addition, eight levels of missing data percentages (independent variable 1) were assessed across each data set type to compare the measures of predictive accuracy, for each of the three missing data theory techniques (independent variable 3). Once the data sets were exported to a flat file in Microsoft Excel, the experiment followed a four-step process, like the research conducted by Idri et al. (2016c). The actual known data values (pretest values) provided the pretest baseline that was used to compare how accurately each missing data theory techniques produced its respective “Predicted Numerical Value” (posttest value). The “Original Numerical Values” (pretest /priori values) were removed-at-random to create missing values within the data matrix by using a random number generator (Idri et al., 2015a, 2015b, 2016b, 2016c; Research Randomizer, 2020). Next, the complete data set generation occurred in which the missing data theory technique

(independent variable 3) treatment variables were then calculated and applied to make each of the 28 data sets complete again. After which, the measurement of predictive accuracy evaluation began, and measured the outcome variables, the error approximation values, by calculating the absolute error and relative error values between the pretest and posttest values from the pre-experiment. ANOVA was used to test the study's null and alternative hypotheses, and to determine if there was a significant interaction between independent variables. This research could mitigate the current gap in literature because it tested if missing data theory techniques improve the reliability and completeness of defense historical data when missing and incomplete values are present in a physical data matrix of a cost estimator using an empirical pre-experimental design.

To further the application of missing data theory to the U.S. defense cost estimation discipline, I modeled the missing data problem by simulating the conditions that defense industry cost estimators, engineering economists, engineering managers, and defense cost estimating repository database administrators experience when they receive a data matrix with missing values. With this pre-experimental study design, I applied three types of missing data theory techniques by administering complete case analysis treatments, single imputation treatments, and multiple imputation treatments on the same group of randomly selected data from a complete U.S. represented cost estimation data matrix. The data matrix contained an appropriate required sample size of DoD software cost estimation programs (see Idri et al., 2015a, 2015b, 2016b, 2016c). The "Predicted Numerical Value", as determined by each missing data theory technique, served as the posttest value in this experiment to help calculate the study's dependent variables, which

were measures of predictive accuracy. Stated differently, the two dependent variables that captured the predictive accuracy for this study were absolute error and relative error. The absolute errors and relative errors were calculated from pretest and posttest values. The study design leveraged a complete data set to allow for the “Predicted Numerical Value(s)” from each data matrix to be assessed against each original “Original Numerical Value(s)” as provided from a nonproprietary data matrix. The data matrix held data sets that were representative of what could be found in databases used by cost estimators, engineering economists, and engineering managers within the defense cost estimating discipline (e.g., from the Functional Academic Cost Analysis Database Environment [FACADE], USASpending.gov [2021], IT Dashboarddata.gov [2021], etc.). This allowed for an empirical examination as to how well missing data theory techniques corrected missing data sets that had missing values. To answer the RQ, I used the pre-experimental research design of the one group pretest-posttest no control group design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Crammer, 2018, Reichardt, 2019; Shadish et al., 2002; Shek & Zhu, 2018; Singleton & Strait, 2010; Thyer, 2012). Significance testing was performed by conducting a two-way repeated measures ANOVA. The *F* ratio was used to test the main effects and interaction between the variables.

This quantitative research method of inquiry was chosen to help determine how well defense cost estimators could handle historical data sets with the use of missing data theory techniques (see Crammer, 2018; Kirk, 2013; Shek & Zhu, 2018; Thyer, 2012). By randomly removing data values from a complete data set, an empirical examination of new data values was quantitatively created and introduced to assess each missing data

theory's ability to improve data quality (see Kirk, 2013). Further details are discussed in Chapter 3, the Methodology section.

Definitions

There are several operational definitions and terms that are unique to both the cost estimation discipline, as well as the missing data theoretical framework. The following terms may have different meanings in other taxonomies and must be defined to understand this research.

Analysis of Variance (ANOVA): “A parametric inferential statistic that examines differences between the means of three more groups in a study, groups exposed to different independent variables (e.g., treatment 1 vs. treatment 2 vs. treatment 3), or longitudinally at least three times for a single group (e.g., pretest, posttest, and at follow-up)” (Thyer, 2012, p. 179).

Business—Cost Estimating: The career field and discipline for the area of business in which “engineering judgment and experience are utilized in the application of scientific principles and techniques to the problems of cost estimation, cost control, and profitability” (Spruill, 2021, p. 2). This U.S. defense career field includes positions that “manage, supervise, lead, or perform scientific work that involves designing, developing, and adapting mathematical, statistical, econometric, and other scientific methods and techniques” (Spruill, 2021, p. 2). In addition, the type of work in this discipline includes “analyzing management problems and providing advice and insight about the probable effects of alternative solutions to these problems” (Spruill, 2021, p. 2).

Cost: Cost is a driving consideration in decisions that determine how systems are developed, produced, and sustained (Garvey et al., 2016).

Cost Analysis: A method of estimating the economic performance of a commodity over its life period (Desai et al., 2016). It is also known as “whole cost accounting” and “total cost of ownership includes estimating all cost from the initial stage through the divestment stage of an investment (Desai et al., 2016, p. 390). Cost analysis is a term that is broadly used to include not only the process of estimating (measuring) the cost of a project but also the process of discovering, understanding, modeling, and evaluating the relevant information necessary to estimate the cost as well as the cost uncertainty and risk (Melese et al., 2015).

Cost estimates: An end-product from cost estimating. It is a critical document needed to request the right amount of budget authority from Congress to fund future investments (Iqbal et al., 2017; Mislick & Nussbaum, 2015).

Cost estimating: Taken from Mislick & Nussbaum (2015, p. 11), “Cost estimating is the process of collecting and analyzing historical data and applying quantitative models, techniques, tools, and databases in order to predict an estimate of the future cost of an item, product, program or task.”

Cost estimation: “The application of the art and the technology of approximating the probable worth (or cost), extent, or character of something based on information available at the time” (Mislick & Nussbaum, 2015, p. 11). Fundamentally, it is a computational process used to predict final project costs (De la Garza & Rouhana, 1995).

It is a specialized function of cost engineers that are concerned with the economic results of an engineering design and modeling (Grimstad et al., 2006; Ostwald, 1974).

Cost estimators/Cost engineers: Practitioners who are cost estimators or system cost engineers who forecast engineering economic requirements. They are responsible for determining the engineering project requirements and resources needed for defense engineering systems and must have access to reliable and complete data sets from historical database repositories or other ad hoc data sources they collect in order to develop accurate engineering economic requirements. Business students, practicing accountants, and economists are closely identified with cost estimating and cost engineering activities (Ostwald, 1974).

Data matrix(singular)/Data matrices(plural): All rows and columns comprised of two or more data sets from different cases (Little & Rubin, 2020).

Data sets: The rectangular column of a data matrix that describes a common set of data from different cases (Little & Rubin, 2020).

Engineering economics: Previously known as engineering economy, engineering economics is the application of economic techniques to the evaluation of design and engineering alternatives. The role of engineering economics is to assess the appropriateness of a given project, estimate its value, and justify it from an engineering standpoint (Farr & Faber, 2018; Fraser & Jewkes, 2013; Newnan et al., 2004).

Experimental design: “A research study in which one or more independent variables are systematically varied by the researcher to determine their effects on

dependent variables” (Thyer, 2012, p. 181). Randomized experiments randomly assign participants to various (a) treatments, (b) control or (c) comparison groups.

Hands-on-treatment-options: The ability to use missing data theory techniques as techniques for the cost estimation discipline to teach cost estimators ways in which they can directly handle unreliable and incomplete data. This includes but is not limited to potential applying missing data theory techniques such as complete case analysis (listwise delete), direct imputation (single or multiple), model-based imputation (full information maximum likelihood), and machine learning methods (García-Laencina et al., 2010).

Interrupted time series (ITS) design: “Longitudinal research in which ongoing repeated measurements of the outcome are made and treatment is introduced at some point, while measurements continue as before” (Thyer, 2012, p. 182).

Missing at random (MAR): Missingness has systematic relationship to observed values, but not missing values (Rubin, 1976).

Missing completely at random (MCAR): Missingness has no systematic relationship to observed or missing values of any variables (Rubin, 1976).

Missing data theory techniques: The different ways in which missing data can be handled. This includes but is not limited to complete case analysis techniques such as the complete case analysis (listwise delete) treatment, direct imputation techniques such as single imputation and multiple imputation, model-based imputation techniques such as full information maximum likelihood and the expectation-maximization (EM) algorithm, and machine learning methods such as ensemble methods, support vectors, and gradient

boosters (García-Laencina., 2010). According to IBM SPSS 25 you can also use the EM algorithm as a single imputation technique as well (IBM knowledge center, 2021).

Missing not at random / Non-ignorable: Missingness has systematic relationship to missing values (Rubin, 1976).

One group pretest-posttest design: “A pre-experimental design involving one group that is pretested, exposed to a form of treatment, and then posttested” (Thyer, 2012, p. 184).

Percentage of missingness: The various percentages in which missing values appear in this study, generally accepted that missing data theory techniques work well at percentages of 40% or lower (Strike et al., 2001).

Predictive accuracy: How close the error approximation is between a data set’s “Original Numerical Value” (pretest/ priori value) as compared to its “Predicted Numerical Value” (posttest/ posteriori value) imputed based on the applied missing data theory technique treatment (Little & Rubin, 2020; Twala et al., 2006).

Pre-experimental design: “A research design that involves studying only a single group of participants, either posttreatment only, or pre- and posttreatment” (Thyer, 2012, p. 184). No control or comparison groups are used (Thyer, 2012).

Quasi-experimental design: “A type of research design in which the treatment and control or comparison groups are not created using random assignment procedures” (Thyer, 2012, p. 185). “It does involve the manipulation of an independent variable and the specification of a test hypothesis” (Thyer, 2012, p. 185).

Removed-at-random-data-values: Values that have been removed at random by using a random number generator (Research Randomizer, 2020).

Software cost estimation: The summation of what is required to build software which includes labor effort, lines of code, and calendar time required to develop, deliver, and maintain any software-based engineering project (Wani et al., 2019).

Assumptions

I assumed that the accuracy of the data used from FACADE was representative of data found in U.S. defense cost estimation discipline based on it being the database used to teach and certify DoD cost estimators and engineers who attend DAU courses. In addition, historical data used from U.S. defense federal public domain databases are assumed to be accurate, and representative of ad hoc data sources that are used by DoD cost estimators and engineers who are practitioners in the Business—Cost Estimating, as well as the engineering economics field of study. Moreover, this pre-experimental study supports the missing data theory mechanism assumption that all data values removed are MCAR and concludes that the missingness of each variable has no correlation to the values of other variables, or to its own known real or ground truth value (Enders, 2010; Rubin, 1976). Lastly, this study supports the assumption that the repository of work breakdown structures and all other historical project documentation used in this study have been collected at the appropriate levels and stored carefully to reflect how actual resources were used to complete past engineering projects. All assumptions were necessary to establish and document prior to empirically testing and measuring each

missing data theory technique on a U.S. defense cost estimation data sets from a data matrix to answer this study's RQ.

Scope and Delimitations

This study was bounded to a pre-experimental design from a public domain data source. The data source is a representative data matrix in which federal U.S. defense cost estimators use as model inputs for software effort estimation to determine what it may cost. All public domain sources for a data matrix below were considered. I was able to have the first item in the list approved for this study. The data matrix used within the DAU BCF 250 Course, a nonproprietary data matrix, received institutional review board (IRB) approval for me to use for my empirical research via IRB approval number 11-13-20-0127578 (Walden University, 2020). A list of all data options that were considered to use were the following, in priority order, to connect with positive social change influence:

1. Functional Academic Cost Assessment Data Enterprise (FACADE) Demonstration and Training Site from the OSD CAPE, as well as the data matrix used in the DAU BCF 250 Course, Software Cost Estimation
2. Software Resources Data Report (SRDR) flat files from Cost Assessment Data Enterprise (CADE) Database
3. Public domain data from IT Dashboard.gov (IT Dashboarddata.gov, 2021)
4. Public domain data on actual DoD spending sites from past years (USAspending.gov, 2021; Federal Procurement Data System – Next Generation, 2021)

The IRB request and approval supported the most valuable social change contribution, the use of the FACADE data matrix that was incorporated into the BCF 250 Software Cost Estimation course was and still is a representative U.S. defense cost estimating data matrix that has been presented to train U.S. defense Business—Cost Estimating students at DAU. Under U.S. copyright law (17 USC§ 105), works created by all federal employees, including DAU, as part of their official duties are in the public domain and may not be copyrighted (2010). This applies not only to printed materials, audiovisual materials, sound recordings, and so forth, but also to content created for the DAU affiliated websites. As a result, this research's findings and results could influence how the discipline's curriculum is taught at the university, closes the gap in literature, and thus incorporates its significance to both contributing to the discipline's practice and social change contribution to improve cost estimators, engineering economists, and engineering managers techniques in software estimation. The missing data theory techniques were tested and applied to nonproprietary U.S. defense cost estimating data, which is the focus of this study's RQ. As a result, general findings and conclusions can be made from this body of work that has a specific focus, and bounded scope.

Limitations

The research design of this study was limited based on the instrumentation selected to test predictive accuracy. I used IBM SPSS 25 as the instrumentation to conduct a pre-experimental design to test the predictive accuracy of missing data theory techniques on a representative U.S. defense cost estimating data matrix. SPSS is recognized in the academic community and has the statistical capability and processing

power to assess data that has incomplete and missing values (Enders, 2010). I leveraged the statistical analysis capability that is provided in the Missing Value Analysis module, a Multiple Imputation functionality of IBM SPSS 25. The Missing Value Analysis module and Multiple Imputation functionality in IBM SPSS 25 has the computational ability to compute traditional missing data theory algorithms. As a result of this functionality, IBM SPSS 25 was applied as the instrumentation for this inaugural study that tested missing data theoretical techniques' predictive accuracy when applied to the U.S. defense cost estimation domain. Despite this being a limitation of this study, treatments were replicated and assessed as a one group pretest-posttest no control group/pre-experimental design intervention.

Not having a control group for the one group pretest-posttest pre-experimental research design was a weakness; however, it was not pertinent for the RQ based on the nature of the group being data vice human beings. For example, in social work, human beings under intervention studies make it difficult to control for outside influences and can skew their responses that may not be isolated, and thus require a control group to compare results (Thyer, 2012). The use of data as the subject in this intervention under a one group pretest-posttest design enabled me to minimize potential threats to internal and external validity because each independent variable completely controlled how I manipulated the pre-experiments in isolation. I controlled the experiments for each data set to only receive three types of treatments, and evaluated them within the confines of this intervention study's independent variables. As a result, I was able to mitigate any confounding or extraneous variables from entering the intervention study, each dependent

variable was instantly evaluated within a short time-box to answer this study's RQ after the intervention.

In addition, the construct of this study remained strong because of its well-defined and focused scope to test and measure the level of predictive accuracy of missing data theory as it pertains to (a) Listwise Deletion (LD) or Complete Case Analysis, (b) Single Imputation and (c) Multiple Imputation on an IRB approved and representative U.S. defense cost estimation data matrix. This narrowed focus is not biased, but it is intentional to address the specific RQ of this study that takes a first look at applying traditional missing data theory to the U.S. defense cost estimation domain, something that has never been done before this intervention study. Further studies can extend the scope of this study and add to the literature to expand outcomes of this analysis.

Significance of the Study

This study is important because “reliable and comprehensive cost data is essential to produce credible cost estimates as required in both (policy) statute and regulation” (Morin, 2017, p. 1). Brown and White (2017) agreed with Morin and reported that the federal defense department lacked the data, both in volume and quality, needed to conduct effective cost estimates. Together, these authors acknowledged that cost estimate realism is essential and needed to support engineering and program managers with the authority to proceed in the development and contractual procurement of critical engineering systems. This study may offer a different perspective on an established problem that historical databases contain substantial amounts of missing data (Strike et al., 2001). Conducting research to “improve analyst productivity, quality of cost

estimates, close data gaps, and provide the cost acquisition, and resource allocation organizations with data required for better analysis and decision-making” could be significant (Morin, 2017, p. 1). The results from this research can be adopted as an option to improve data quality, improve analyst productivity, and minimize the unreliable and incomplete data problem experienced by cost estimators, engineering economists, and the engineering managers that rely on what is taught within the Business—Cost Estimating body of knowledge.

Significance to Theory

The outcome of this study may offer defense industry cost estimators, engineering economists, engineering managers, defense cost estimating repository database administrators, and possibly data scientists with an objective option in how to deal with missing, incomplete, or unreliable data values when they appear within a data matrix. Applying and testing missing data theory on an actual complete data set that is relevant to the problem could provide the empirical evidence needed to prove or disprove how well various missing data theories are able to fill missing data value gaps. Contingent on the outcomes observed after randomly removing variables to simulate a missing data problem, this could improve the missing, incomplete, and unreliable data problem that is experienced within the U.S. defense cost estimation discipline. In addition, U.S. defense cost estimators tend to build models with small data matrices, N less than or equal to 30, in which an empirical study that tested the performance of small sample size data sets, and how well missing data theories’ predictive accuracy levels were explored.

Significance to Practice

Cost estimators of defense weapon systems must have access to reliable and complete data sets from the historical database repositories and other sources they access to develop accurate engineering economic requirements. Cost estimates, the end-product from cost estimating, is a critical document needed to request the right amount of budget authority from Congress to fund any future investments (Mislick & Nussbaum, 2015). When databases have null values, obvious errors, and blank cells because of various systemic data problems, it is up to the cost estimator to make the decision as to how to use this type of data value within a data matrix to feed a cost estimate element. In layman's terms, there is no standard approach taught to defense cost estimators in what data values to use or not use in their physical data matrix when the missing, incomplete, or unreliable data values appears (DAU, 2018a). With over 250 defense cost estimators within the Business—Cost Estimating career field, there is no established standard as to how to handle this problem within the defense cost estimating discipline (DAU, 2018a, 2018b). Offering engineering managers and cost estimators within the discipline additional options to determine how to handle missing, incomplete, or unreliable data values, could reduce the number of flawed cost estimates that lead to program cost overruns and unplanned additional federal budget request (Schwartz & O'Connor, 2016).

Significance to Social Change

Accurately forecasting estimates for engineering requirements could save projects and programs from growing cost overruns and improve U.S. federal planning decisions (Christensen, 1993; Christensen & Gordon, 1998; Deloitte, 2016; Saeed et al., 2018). In

addition, positive social change could be realized by improving the current techniques cost estimators and engineering managers use to produce and provide more accurate, reliable, and credible cost estimates to federal decision makers (Govinfo, 2020).

Moreover, research that could advance cost data quality and improvement efforts could also increase the amount of historical DoD cost data that can be used in analyses. Overall, a new way of doing business may save cost estimator's, engineering economists', engineering manager's and database administrator's valuable time by using a newly proven technique to improve data in a shorter amount of time. In turn, this contribution to the cost estimation discipline has the potential to reduce the cost of an estimator's research time and reduce the cost required to collect additional data.

Summary and Transition

As a starting point, Chapter 1 contains the problem and purpose statement of this empirical study and addresses the gap in the DoD cost estimation discipline current literature to support future improvements in both the engineering economics and engineering management fields of study. This chapter also contains the specific research goal and objectives, and the scope of this research. Furthermore, this chapter also describes the motivation of this research project to improve the state of practice and bring about social change.

Chapter 2 contains the theoretical framework of missing data theory that grounds this body of research, followed by the literature research strategy. I provide an overview of data quality requirements that has been levied on the Business—Cost Estimating discipline through U.S. policy statutes and regulations for reliable and complete cost

estimation and cost analysis data. After which, I describe the issues experienced by cost estimators, engineering economists, and engineering managers in DoD cost estimation and analysis fields and highlights the gap that this research addresses by discussing a topic that has been silent within the U.S. defense cost estimation discipline since its inception, circa 1972. Next, I introduce how other disciplines have empirically researched and used the missing data theoretical framework and its techniques as a tool to handle their unreliable and incomplete data problems and needs. Finally, I describe the contribution of this body of work: conducting empirical research to determine which missing data theory technique(s) best lends itself to improving predictive accuracy when applied to a U.S. defense cost estimating matrix. Stated comprehensively, the full purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. This research specifically narrows in on the area of software cost estimation which is predominately discussed and supported in the literature as an area that cost overruns frequently exist, as well as has more conversations occurring in scholarly peer reviewed journal articles, conference proceedings, and academic well-renowned books in cost estimation (see Boehm, 1981; Idri et al., 2016c; Jones, 2007; Jing et al., 2016; Strike et al., 2001).

Chapter 3 contains the general view and detailed view of the empirical research methodology to answer this study's RQ. This chapter fully describes the pre-experimental design selected to investigate a representative U.S. defense cost estimation data matrix in the public domain in order to empirically deal with missing values and outliers when used to build cost estimation relationships and other forecast that require reliable and complete data for cost analysis. Chapter 3 includes the research design method, theoretical method of inquiry, justification of the research method, the justification of the intended sample and sample size, method of data collection and procedures, data management, data analysis technique and research method, issues of ethical considerations, reliability and validity, and instrumentation.

Chapter 4 contains the results of the final study. It includes describing the data collection that occurred and highlights new discoveries identified because of executing the three applied missing data theory techniques. Most importantly, all pre-experimental study results from this one group pretest-posttest no control group/pre-experimental design for 4,704 ($3 \times 56 \times 28$) treatments were recorded and can be found in the Appendix A. Summary tables provide descriptive statistics that appropriately characterize the starting purposive sample of 30 out of 50 analogous and synthetic software development programs that were then randomly sampled to create the artificially induced missing data problem that required 56 missing data treatments per data set. Chapter 4 includes interesting findings and results which includes the statistical assumptions used to answer this study's RQ and hypotheses, including exact statistics and associated probability values and post-hoc analyses of statistical tests referred to that can be found in Appendix

C. All results are accurately presented and are aligned to the RQ and study's hypotheses, design, and analysis.

Lastly, the study results and outcomes from Chapter 4 are interpreted into the conclusion drawn in Chapter 5. Chapter 5 also includes the recommendations and further studies that could be continued because of this research. The conclusions, limitations, and recommendations are clearly described for the scope of this study and can now be integrated into the state of knowledge described in the literature review to close a gap in the Business–Cost Estimating discipline.

Chapter 2: Literature Review

With DoD cost overruns rising, U.S. defense cost estimators need more options available to them to know how to handle the unreliable and incomplete data they use to build estimates, which allows them to forecast life-cycle cost analysis for a myriad of engineering-based acquisition projects and programs (DAU, 2018a, 2018b; GAO, 2009, 2020; Morin, 2017). The specific management problem is that there is a lack of research into the techniques to handle the unreliable and incomplete data problem. Consequently, the research problem is that there is a lack of knowledge and understanding among cost estimators about what options they have to improve the data quality of data sets with limited, incomplete, or unreliable data, which prevents them from forecasting accurately the life-cycle costs for a myriad of engineering-based acquisition projects and programs (DAU, 2018a, 2018b; GAO, 2009, 2020; Morin, 2017). In a government publication memo, Morin's (2017) approach to the problem was to start a data collection effort through updating eight topics within two major policies to improve data quality and estimation conditions (Department of Defense Instruction [DoDI], 2017; Department of Defense Manual [DoDM], 2011; Morin, 2017). This approach supports that research into cost analysis data quality is significant, and that improvements are still needed.

Unfortunately, changing policy to create better data collection efforts only looks at one aspect of the problem but fails to address how cost estimators could handle the missing data problem when they have physical historical data sets in front of them that are missing and incomplete. Through the lens of missing data theory, several empirical researchers have addressed the needs of both social and natural scientists across many

disciplines with options to deal with handling their data matrices that may have unreliable, incomplete, or even completely missing values (see Aittokallio, 2009, Baraldi & Enders, 2010; DeLeeuw, 2001; García-Laencina et al., 2010, 2013; Tsiriktsis, 2005). Moreover, many empirical researchers have assessed predictive accuracy on data matrices and have conducted experimental designs using missing data theory (Lin & Tsai, 2019). Unfortunately, none have been applied to any U.S. defense software cost estimation data matrices (Khoshgoftaar & Van Hulse, 2008; Song et al. 2008; Van Hulse & Khoshgoftaar, 2014). Currently, there is a gap that needs to be addressed within the literature of the U.S. defense cost estimation discipline that describes how defense cost estimators could handle its physical unreliable and incomplete data problem when historical data sets have missing values (Brown & White, 2017; DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013; GAO, 1972, 2009, 2020; Mislick & Nussbaum, 2015). The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages.

Chapter 2 contains the theoretical framework of missing data theory that grounds this body of research, followed by the literature research strategy. I then provide an overview of data quality requirements that have been levied on the Business—Cost Estimating discipline through U.S. policy statutes and regulations for reliable and

complete cost estimation and cost analysis data. After which, I describe the issues experienced by cost estimators, engineering economists, and engineering managers in DoD cost estimation and analysis fields. Within this section, I highlight the gap that this research study addresses by discussing a topic that has been silent within the U.S. defense cost estimation discipline or curriculum since its inception circa 1972 (see 10 U.S. Code § 1746, 2012; GAO, 1972, 2009). In addition, this topic has not been included in Business—Cost Estimating curriculum at DAU which began in the 1990s (10 U.S. Code § 1746, 2012). Next, I introduce how other disciplines have empirically researched and used the missing data theoretical framework and its statistical and machine learning techniques as a tool to handle their unreliable and incomplete data problems (see Ghorbani, & Desmarais, 2017). Finally, I describe the gap in the literature: the lack and need for empirical research that could determine which missing data theory technique(s) best lends itself to determine what the predictive accuracy of missing data theory techniques are when applied to U.S. defense cost estimating matrices. This research specifically focuses on the area of software cost estimation which is predominately discussed and supported in the literature as an area that cost overruns frequently (see Jones, 2007; Strike et al., 2001).

Literature Search Strategy

The following section is a review of the literature for the research study and question: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix?

This review of the literature provides a scholarly analysis of government documents, government websites, conference proceedings, scholarly peer-reviewed articles, and books within the systems engineering and engineering economics subdiscipline of defense agency software cost estimation training practices. Furthermore, the search supports why applying the theoretical framework of missing data theory should be tested on U.S. defense cost estimation data to fulfil this current gap in the literature. With the cost estimation discipline being faced with data reliability and completeness challenges within the DoD, this study is narrowly focused on how unreliable and incomplete data matrices are handled in U.S. defense cost estimation data matrices that are software lines of code and effort hour based.

Literature found specifically between 2015-2020 was used to describe the current environment and scholarly review around the U.S. defense software cost estimation, unreliable data, and incomplete data problem. Literature surrounding the seminal theory of missing data and various statistical and machine learning techniques is also referenced from the literature and covers the 1976-2020 timeframe. The total number of references in the literature review is 142, of which 10% are from seminal theorists, 67% are from scholarly peer-reviewed sources and 60% were published within 6 years of my expected graduation. References include a full range collection of materials for this capstone topic that cites seminal theorists, government documents, government websites, conference proceedings, scholarly peer-reviewed articles, books, and one technical report.

In starting this research, I began by determining what key words were tied to this capstone research. The following key words and Boolean Strings were initially used to search my topic of interest:

1. missing data OR unreliable data
2. predictab* OR imputation OR theor* OR experimental
3. (Miss* OR incomplet*) AND (value OR attribute OR data* OR input OR variable OR feature) AND (experiment* OR metric OR measur* OR assess* OR evaluat* OR predict*) AND (software OR application OR program OR system) AND (Engineering OR maintenance OR science OR develop* OR test* OR construct* OR design* OR project OR effort OR cost OR requirement OR quality OR process) AND (imput* OR deal*OR handl*)
4. “unreliable data” OR “incomplete data” AND “software cost estimation”
5. "software cost estimation" AND "missing data" AND "empirical"
6. "software cost estimation" AND "missing value" AND "empirical"
7. "software cost estimation" AND "missing value" AND "experimental"
8. "software cost estimation" AND "missing value" AND "experimental design"

Based on the articles that have informed me on this area of research, the constructs of why this study was pursued was indeed informed via this literature review. Since this is the first body of research in respect to empirically testing U.S. defense cost estimation data, I focused on the traditional missing data theory as a logical entry point, vice charting into advanced missing data techniques to evaluate to support the U.S. defense cost estimation discipline’s unreliable data problem. I also chose to address this

body of work with a narrower focus on software cost estimation because the majority of scholarly literature discussed falls into the software project and measurement subcategory when discussing the topic of cost estimation (see Abnan & Idri, 2018; Huang et al., 2015a, 2015b, 2017; Idri et al., 2016a, 2016c; Soltanveis & Alizadeh, 2016; Strike et al., 2001; Twala, 2017).

Theoretical Foundation

Beginning in the mid-1970s, seminal works in missing data theory began to appear, and established principles that have been applied to the missing data problems in respect to survey and observed data housed in databases (Little & Rubin, 1976; Rubin, 1976). The main premise behind Rubin's (1976) theory work was that missingness was a variable that had a probability distribution around it which brought a new construct to think about missing data. Applied missing data theory, to include its statistical and machine learning techniques, are commonly used within the literature of various fields as an option to replace missing data values with substitution values (Aittokallio, 2009; Garciaarena, & Santana, 2017). Standard statistical methods are used to assess and analyze rectangular data matrices in which rows of the data matrix represent units, and the columns represent characteristics of each unit (Little & Rubin, 2020). The entries are typically numerical, and are continuous variables such as age or income, or categorical variables such as grade or gender (Little & Rubin, 2020). The major theoretical proposition is that through statistical analysis with missing data, an analyst could effectively predict or impute an unobserved value can add meaning to a data matrix (Little & Rubin, 2020). With a theoretical construct that proposes to effectively predict

unobserved value, testing and measuring its predictive accuracy through an evidenced-based approach would be useful.

To apply the theory into practice appropriately, one must make assumptions on the missing data mechanism (Little & Rubin, 2020). Rubin's (1976) principle to define the missing data mechanism as MAR, MCAR, or MNAR/NI is important to understand because it allows a researcher to perform a proper treatment to address a data matrices' missingness. The mechanisms describe the bias the missing data exerts on a missing data analysis in which the true goal is to minimize bias with unbiased parameter estimates (Rubin, 1987). Rubin's (1976) missing data mechanisms are essentially the assumption that govern the performance of the analytic technique based on the property of the missing data. The properties of missing data inform analysts on the relationship between the propensity of the data that is missing, and the following (Rubin, 1987):

- The variable with the missing data
- The other variables of fully observed data
- The hypothetical mechanism chosen based on the underlying missing data as MAR, MCAR, or MNAR/NI.

Properly applying this theory requires an analyst or researcher to understand how their data was acquired in order to make the right assumptions, and select the mechanism that can support the data matrix that has missing values.

Pre-2015 literature was comprised of roughly four key bodies of work that provided evidence that data quality issues in software estimation historical sets were leveraging missing data theory techniques. Research by Strike et al. (2001) discussed the

difficulties in historical databases used for software cost estimation and tested the performance of listwise deletion, mean imputation, and eight different hot deck imputation methods. In addition, Myrtveit et al. (2001) compared four missing data techniques (MDT) in the context of ERP software cost modeling and evaluated list wise deletion (LD), mean imputation (MI), similar response pattern imputation (SRPI), and full information maximum likelihood (FIML) using the International Software Benchmarking Standards Group (ISBSG) database. Applying missing data theory to improve data reliability and completeness was found within the literature and other researchers evaluated techniques to improve software estimation data quality issues.

Moreover, Cartwright et al. (2003) examined the quality of fit of effort models derived by stepwise regression by comparing raw data sets with values that were imputed by various techniques. From the comparison, Cartwright et al. (2003) found that k-nearest neighbor (k-NN) and sample mean imputation (SMI) significantly improved the model fit, with k-NN giving the best results in the data sets. In addition, research by Sentas and Angelis (2006) investigated and suggested imputation using multinomial logistic regression (MLR) and applied it to projects in the ISBSG software database. This study also compared MLR to other techniques of handling missing data to include listwise deletion (LD), mean imputation (MI), expectation maximization (EM) and regression imputation (RI) under different patterns. In summary, several non-U.S. defense cost estimating disciplines dealt with trying to solve its unreliable, incomplete, and missing data problems, akin to the interest of this research.

The conversation in the literature on studying missing data theory took a pause but started to resurface in software engineering and estimation. Idri et al. (2015a; 2015b) conducted a systematic mapping study of missing values techniques in software engineering data to explore how research was conducted within the discipline. The following year, Idri et al. (2016c) determined that missing data is a widespread problem based on their earlier work, and investigated specifically analogy-based software development estimation and evaluated the predictive performance power of toleration, deletion, and k-nearest neighbor (KNN) imputation using Euclidean distance and Manhattan distance techniques by conducting 1,512 experiments on seven data sets. Jing et al. (2016) conducted seven experiments and proposed the use of low-rank recovery semisupervised regression (LRSR) imputation as a better method than other imputation methods they compared. Moreover, research by Twala (2017) investigated a new probabilistic supervised learning approach that incorporates missingness to improve software effort development predictive accuracy. Abanane and Idri (2018) evaluated four missing data theory techniques using four mixed data sets. Lastly, research from Majeed (2018) investigated how to develop model-based estimation approaches and applied them to the missing data problem as well. With these more recent research efforts surrounding experimenting with data to better understand missing data theory techniques and their utility, extending this type of research through inquiry can extend the overall state of knowledge for this area.

Disciplines that have leveraged missing data theory can be found in the social sciences and the physical science for both research, surveys, databases, and other applied

purposes (Allison, 2000, 2002, 2010; Blankers et al., 2010; Enders, 2010; Little & Rubin, 2014). In addition, experimental designs to test and measure predictive accuracy, similar to this study, have also been conducted on both simulated and historical software data matrices derived from the ISBSG, China Software Benchmarking Standards Group, and University of California at Irvine database repositories (azzahra Amazal et al., 2014; González-Ladrón-de-Guevara et al., 2016; Jeffery et al., 2000; Khoshgoftaar & Van Hulse, 2008; Song et al., 2008; Van Hulse & Khoshgoftaar, 2014; Zhang et al., 2011). This study extends this knowledge by testing and measuring 30 out of 50 analogous and synthetic software development programs from the U.S. defense cost estimation discipline.

Missing data theory was chosen because literature has shown it to be a feasible alternative to improve data quality in many examples from the literature (Allison, 2002; Graham, 2012; Horton & Kleinman, 2007; Jadhav et al., 2019; Myrtveit et al., 2001; Schafer, 1997; Schafer & Graham, 2002). The missing data theoretical framework relates to the present study since it is the RQ that is being challenged by testing its predictive accuracy when applied to U.S. defense cost estimating nonproprietary software program data. This study will challenge as well as build upon the existing theory in respect to this study's evaluation and results to test and determine if missing data theory serves as a feasible alternative to improve the data quality of the U.S. defense cost estimation unreliable and incomplete data matrix problem. The intent of this study was to ascertain how effective missing data theory's predictive accuracy is when applied to the defense cost estimating discipline's variant of software effort missing values when data

preprocessing for estimation. Since this theory addresses data completeness, it was most appropriate to use since several years of research have validated that the theory provides effective techniques to predict and fill in data gaps in many disciplines (Graham, 2009, 2012; Strike et al., 2001).

This theory is a starting point to determine which missing data theory technique could improve the state of the defense cost estimating data reliability and completeness problem. This framework allowed me to look at missing data through a narrower lens by testing various traditional missing data theory techniques (e.g., complete case analysis, imputation, single imputation, and multiple imputation) to inform and update how one could handle and treat areas of missing data within the defense cost estimation discipline (Idri et al., 2016b; Myrtveit et al., 2001a; Strike et al., 2001). Furthermore, this research could inform how this topic will be addressed within academic books and journals involving engineering economics/management, software engineering economics, financial decision sciences, and business analytics, and cost estimation (Boehm, 1981; DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013).

Literature Review

From the perspective of U.S. public policy statutes and regulations, the U.S. defense cost estimating discipline, and the current Business—Cost Estimating curriculum at Defense Acquisition University (DAU), there is a lack of instruction in which cost estimators, engineering economists, and engineering managers can apply to handle the unreliable and incomplete engineering project data matrix problem they face (DAU, 2018a; GAO, 2009, 2020; ICEAA, 2019). According to a U.S. defense based Joint

Agency CER Handbook (2018), “data sets with missing and incomplete data” is a data analysis challenge and states that the “best course of action is to first attempt to remedy the problem by collecting more data, finding the information from the collected data set, and determining the cause of the unusual observations, respectively” (p. 221). This government document also acknowledges that it is “not always possible to correct such errors” and that it is important for estimators to understand the implications of these challenges, and to proceed with their analysis under caution (Joint CER Handbook, 2018, p. 221). The literature does not inform how cost estimators who leverage the defense Business—Cost Estimating curriculum at DAU directly handle the unreliable and incomplete engineering project data matrix problem other than through recognizing the problem through defense government documents and making changes to public policy (GAO, 2020).

U.S. Public Policy Requires Data Reliability/Completeness in Cost Estimation

There are over 50 federal public policy documents that apply today to the Business—Cost Estimating discipline and requires the production of reliable cost estimates (DAU, 2018b). Salient laws, statutes, regulations, policies, guidance, directives, and even manuals are sources of criteria that are currently available to U.S. defense cost estimators that inform how they develop their cost estimates. Most notably, Title 10 U.S. Code § 2334 (2017) is very clear in its expectations and provides the following law that states the U.S. DoD Armed Forces must:

“ensure that cost estimates are developed, to the extent practicable, based on historical actual cost information that is based on demonstrated contractor and

Government performance and that such estimates provide a high degree of confidence that the program or subprogram can be completed without the need for significant adjustment to program budgets”.

This General Military Law under Chapter 137, Part IV, Service, Supply and Procurement, acknowledges that cost estimates forecast engineering project and program budgets. This law acknowledges that historical actual cost information is expected as a matter of law for cost estimation developments but recognizes that this is not always practical.

U.S. public policy continuously gets updated within various government documents in order to provide the *Business-Cost Estimating* discipline and the U.S. cost estimator population with the “processes, procedures, and tools”, as well as legal backing to support the requirement to produce reliable and complete estimates (GAO, 2009, 2020, p. 3; Morin, 2017). These public policy documents are vital and inform the discipline about the “criteria” cost estimators must follow as they go about forecasting life-cycle cost for a myriad of engineering-based acquisition projects and programs (DoD, 2020; GAO, 2009, p. 25). Table 1 below provides a select list of federal and DoD public policy in order to highlight the breadth of government documents that currently supports U.S. defense cost estimator’s data reliability and completeness requirement in cost estimation (DAU, 2018b; GAO, 2009, p. 26-29, GAO, 2020).

Table 1

Select U.S. Public Policy Criteria and Requirements to Produce Reliable and Complete Cost Estimates

Id	Business—Cost Estimating policy	Type	Current publication	Original publication
1	DoDI 5000.73 Cost Analysis Guidance and Procedures	Guidance	2020	2006
2	DoDI 5000.02 Operation of the Adaptive Acquisition Framework (AAF)	Guidance	2020	
3	Army Cost Analysis Manual	Manual	2020	
4	DoDI 5000.74, “Defense Acquisition of Services	Policy	2020	
5	OMB Circular A-11, Part 7 - Preparation, Submission, and Execution of the Budget	Policy	2019	2006
6	10 U.S. Code § 2334 - Independent cost estimation and cost analysis	Legislation	2019	2017
7	SECNAVIST 7110.12, Department of the Navy	Policy	2019	

	Acquisition Program Cost			
	Analysis			
8	MIL-STD-881-D Work	Standards	2018	
	Breakdown Structure			
9	Joint Agency Cost Estimation	Handbook	2018	
	Relationships (CER) Handbook			
10	DoDI 5000.02T Operation of the	Policy	2017	2015
	Defense Acquisition System			
11	DoDD 7041.03 CE-01	Policy	2017	1995
	Economic Analysis for			
	Decision-making			
12	DoDD 7045.14, The Planning,	Directive	2017	2013
	Programming, Budgeting, and			
	Execution (PPBE) Process			
13	DoDI 5000.75 “Business	Policy	2017	
	Systems Requirements and			
	Acquisitions”			
14	Directive-type Memorandum	Directive	2017	
	(DTM) 17-001: Cybersecurity			
	in the Defense Acquisition			
	System			

- | | | | |
|----|--|-------------|------|
| 15 | SECNAVINST 5000.42

Department of the Navy

Accelerated Acquisition for the

Rapid Development,

Demonstration and Fielding of

Capability | Policy | 2016 |
| 16 | National Defense Authorization

Act (NDAA), Section 804,

“Middle-tier Acquisition for

Rapid Prototyping and Rapid

Fielding” | Legislation | 2016 |
| 17 | Office of the Secretary of

Defense (OSD) Cost

Assessment and Program

Evaluation (CAPE), “Inflation

and Escalation Best Practices for

Cost Analysis” | Guidance | 2016 |
| 18 | Implementation Directive for

Better Buying Power 3.0 -

Achieving Dominant

Capabilities through Technical

Excellence and Innovation | Directive | 2015 |

19	Defense Federal Acquisition Regulation Supplement, 234.7101 Cost and Software Data Reporting (CSDR)	Regulation	2014
20	DoDI 7600.02, Audit Policies	Policy	2014
21	Office of the Secretary of Defense (OSD) Cost Assessment and Program Evaluation (CAPE), Operating and Support Cost Estimating Guide	Guidance	2014
22	Office of Management and Budget (OMB), “Improving Information Technology (IT) Project Planning and Execution,” Memorandum for Chief Information Officers	Policy	2014
23	Joint Agency Cost Schedule Risk and Uncertainty Handbook	Handbook	2014
24	AF Policy Directive 65-5, Cost and Economics	Policy	2013

25	DoDI 5010.40 Managers' Internal Control Program Procedures	Policy	2013
26	DoDD 2140.02, Recoupment of Nonrecurring Costs (NCs) on Sales of U.S. Items	Directive	2013
27	Implementation Directive for Better Buying Power 2.0 - Achieving Greater Efficiency and Productivity in Defense Spending	Directive	2013
28	Independent Cost Estimates; Operational Manpower Requirements, 10 U.S.C. § 2434	Legislation	2012
29	DoD Directive 5105.84, “Director of Cost Assessment and Program Evaluation (DCAPE)”	Directive	2012
30	DoDM 5000.04-M-1, Cost and Software Data Reporting (CSDR) Manual	Manual	2011

31	Government Performance and Results Act (GPRA) Modernization Act of 2010, Pub. L. No. 111-325, 124 Stat. 3866	Legislation	2011	2010
32	Better Buying Power: Guidance for Obtaining Greater Efficiency and Productivity in Defense Spending	Guidance Memo	2010	
33	Interim Acquisition Guidance for Defense Business Systems (DBS)	Guidance	2010	
34	Weapon Systems Acquisition Reform Act of 2009, as amended	Legislation	2009	
35	National Security Space Acquisition Policy DoD Interim Guidance	Guidance	2009	2004
36	SAR: Selected Acquisition Reports, 10 U.S.C. § 2432	Legislation	2006	1968
37	Unit Cost Reports (“Nunn-McCurdy”), 10 U.S.C. § 2433	Legislation	2006	1982

- | | | | |
|----|---|-------------|------|
| 38 | Major Automated Information System Programs, 10 U.S.C. §§ 2445a–2445d | Legislation | 2006 |
| 39 | Clinger-Cohen Act of 1996, 40 U.S.C. §§ 11101–11704 | Legislation | 1996 |
-

In a government publication memorandum entitled DoD Cost Analysis Data Improvement, Morin’s approach to the problem was to start a data collection effort through updating two policies to improve eight topic areas to include data quality and estimation conditions (Morin, 2017). This approach supports that research into cost analysis data quality is significant, and that improvements are still needed. There is a current gap that needs to be addressed within the literature of the U.S. defense cost estimation body of knowledge that describes how defense cost estimators should handle its unreliable and incomplete data problem when historical data sets have missing values (DAU, 2018a; GAO, 1972, 2009, 2020; Mislick & Nussbaum, 2015;). In the past, GAO reported that the federal defense department lacked the data, both in volume and quality, needed to conduct effective cost estimates (Brown & White, 2017). Chapter 2, DoD 5000.4-M identifies four major analytical methods or cost estimating techniques used to develop cost estimates for engineering and acquisition programs: a) analogy, b) parametric (statistical), c) engineering (bottoms up) and d) actual costs (DoDI, 2017; DoDM, 2011; Williams & Barber, 2011). With over 250 defense cost estimators within

the Business—Cost Estimating career field, there is no established standard as to how to handle this problem within the defense cost estimating discipline (DAU, 2018a).

The ability to generate reliable cost estimates is a critical function that supports the Office of Management and Budget’s (OMB) capital planning process that cannot be ignored when major engineering projects and programs are needed to support the U.S. DoD (GAO, 2009, p. i). The capital planning process, prescribed through the OMB Circular A-11 regulation, is required for all U.S. defense and federal agencies to adhere to as they forecast cost in their annual budget justification and submissions that supports the creation of the U.S. annual federal budget (GAO, 2009, p. i).

Effect of U.S. Public Policy on Cost Estimation Data Reliability/Completeness

A longstanding problem in the U.S. defense cost estimating discipline is that many cost estimators cannot generate reliable cost estimates to support the U.S. defense and federal budgets because the underlying and historical data available to them to generate accurate estimates is incomplete or missing (Morin, 2017; GAO, 2020). As far back as 1972, a Government Accountability Office (GAO) reported a damaging finding in an assessment of U.S. defense cost estimates that “known costs had been excluded without adequate or valid justification” (p. 1). Within the same assessment, the GAO (1972) also commented that, “historical cost data used for computing estimates were sometimes invalid, unreliable, or unrepresentative” (p. 1). Thirty-seven years later, another GAO report stated the same problem, and attempted to provide additional guidance and structure for cost estimators to use more reliable data (2009) as a fix.

The lack of reliable and comprehensive data available in the defense industry has contributed greatly to the fact that managers and engineers are unable to estimate project and program requirements accurately, thus producing inaccurate economic forecast for a very long time (GAO, 1972, 2009; Jorgensen, 2006, Morin, 2017). To better understand the impact of inaccurate economic forecasting, one needs to understand what a cost underrun or overrun is. Cost underruns and overruns are a metric that measures forecasted cost estimates of schedule, budget, and manpower requirements compared to what is ultimately spent on an engineering project or program (Saeed et al., 2018).

Unfortunately, reported cost overruns within the DoD major defense acquisition program (MDAP) portfolio programs in 2015 was \$468 billion, up from \$295 billion in 2008 (Deloitte, 2016). Other studies have suggested that individual DoD programs have cost overruns as high as 40% and are projected to overrun closer to 51% by the year 2020 (Christensen, 1993; Deloitte, 2016). In 2020, the U.S. Treasury Department reported that total defense budget plans will cost the country over \$718 billion dollars, a \$33 billion or 5% increase from 2019 enacted levels (U.S. Government, 2020). At such high spending levels, solving any potential cost overruns and mishaps due to unreliable and incomplete data is needed. Currently, there is a gap within the literature of U.S. federal defense cost estimating body of knowledge as to how cost estimators should handle its unreliable data problem which can include having missing, incomplete, and erroneous data (Brown & White, 2017; GAO, 1972, 2009, 2020).

The general management problem is that despite this, cost estimators do not always have reliable and complete data sets to use when they attempt to forecast life-

cycle costs for a myriad of engineering-based acquisition projects and programs and may sometimes forecast costs inaccurately that engineering managers depend on (GAO, 2009, 2020; Jorgensen, 2006; Morin, 2017). In 2015, cost estimate growth was reported as cost overruns within the DoD's Major Defense Acquisition Programs at \$468 billion, up from \$295 billion in 2008 (Deloitte, 2015). Other studies state that individual DoD engineering-based acquisition projects and programs experienced cost overruns as high as 40% and were projected to overrun closer to 51% by the year 2020 (Christensen, 1993; Deloitte, 2015). These costs overrun statistics support the currency and relevancy that this problem must be addressed. This problem is significant to the multi-disciplines of cost estimation which spans business, engineering economics, and systems engineering disciplines (DAU, 2018a; Farr & Faber, 2018; Fraser & Jewkes, 2013; Parnell, 2017).

Gap: Expanding Options to Handle the Estimator's Unreliable/Incomplete Data Problem

A cost estimate is an evaluation and analysis of future costs of hardware, software and/or services (Mislick & Nussbaum, 2015; Williams & Barber, 2011). Cost estimates are generally derived from historical cost, performance, schedule, and technical data associated with similar items or services (Mislick & Nussbaum, 2015). In general, the cost estimating technique used by cost estimators to forecast future life cycle cost for an acquisition program progresses from the analogy to actual cost method as that program becomes more mature and more information is known (Williams & Barber, 2011). The analogy method is most appropriate early in the program life cycle when the system is not yet fully defined. (DoDI, 2017; DoDM, 2011; Williams & Barber, 2011). This

assumes there are analogous systems available for comparative evaluation (Williams & Barber, 2011). As systems begin to be more defined when the program enters a more mature phase of engineering & manufacturing development (EMD) (DoDI, 2017; DoDM, 2011; Williams & Barber, 2011). Estimators are then able to apply parametric once physical manufacturing occurs and actual data is produced for an estimator to use in cost estimation (Williams & Barber, 2011). Estimating via engineering build-up tends to begin in the latter stages of EMD and low-rate initial production (LRIP) when the design is fixed, and more detailed technical and cost data are available (DoDI, 2017; DoDM, 2011; Williams & Barber, 2011). Once the system is being produced or constructed (i.e., LRIP and Full Rate Production), the actual cost method can be applied as a cost estimation technique (Williams & Barber, 2011).

In April 2016, a government document was issued that stated that the Business – Cost Estimating career field had competency gaps that were identified in a consolidated survey and assessment comprised of formal representatives from all U.S. Defense Military Services and the DoD 4th Estate (Burke & Spruill, 2016). Their assessment concluded that a “training gap exists in software cost estimating”, and stated that a new course would be developed, and thus added this course as a new certification requirement (Burke & Spruill, 2016). This same government document also acknowledged that addressing this gap aligned to improving the professionalism of the U.S. defense acquisition workforce, to include cost estimators (Burke & Spruill, 2016). Even more so, assessing the Business—Cost Estimating competency and mitigating a newfound gap by incorporating software cost estimation training supported the Under Secretary of Defense

(Acquisition, Technology, and Logistics) (USD (AT&L) Ashton Carter’s (2010) and Frank Kendall’s (2013) Better Buying Power initiatives to “obtain greater efficiency and productivity in defense spending” (p. 1; Burke & Spruill, 2016; Carter, 2010; Kendall, 2013, 2015). Ironically, historical data sets used in software cost estimation are known to have missing values and have been stated by several authors on this topic (Brown & White, 2017; Jing et al., 2016; Strike et al., 2001). Table 2 below provides a summary of what historical software cost estimation data fields were included in the BCF 250 Software Cost Estimation course for student to use as a synthetic U.S. representative software cost estimation data matrix. (DAU, 2018b).

Table 2

Summary from Data Used within a Synthetic and Complete U.S. Representative Software Cost Estimation Data Matrix Used at Defense Acquisition University (DAU)

Id	Data Sets in Matrix	Data Type Description	Number of Cases
1	Software Intensive Program	Nominal (Synthetic DoD MAIS Program Unique Names)	50
2	Mapped Application Type	Nominal (Dummy Numerical Variables 0,1,2, etc.)	50
3	Operating Environment	Nominal (Dummy Numerical Variables 0,1,2, etc.)	50
4	Primary Programming Language	Nominal (Dummy Numerical Variables 0,1,2, etc.)	50

5	Development Paradigm	Nominal (Dummy Numerical Variables 0,1,2, etc.)	50
6	Upgrade/New	Nominal (Dummy Numerical Variables 0,1,2, etc.)	50
7	Number of External Interface Requirements	Numerical	50
8	Initial SLOC – New	Numerical	50
9	Initial SLOC – Modified	Numerical	50
10	Initial SLOC – Reused	Numerical	50
11	Final SLOC – New	Numerical	50
12	Final SLOC – Modified	Numerical	50
13	Final SLOC – Reused	Numerical	50
14	DM % - Modified*	Numerical	50
15	CM % - Modified*	Numerical	50
16	IM % - Modified*	Numerical	50
17	DM % - Reused*	Numerical	50
18	CM % - Reused*	Numerical	50
19	IM % - Reused*	Numerical	50
20	Final Software Requirements Analysis Effort Hours	Numerical	50

21	Final Software Architectural Design Effort Hours	Numerical	50
22	Final Software Detailed Design Effort Hours	Numerical	50
23	Final Software Construction Effort Hours	Numerical	50
24	Final Software Integration Effort Hours	Numerical	50
25	Final Software Qualification Testing Effort Hours	Numerical	50
26	Final Software Documentation Management Effort Hours	Numerical	50
27	Final Software Configuration Management Effort Hours	Numerical	50
28	Final Software Quality Assurance Effort Hours	Numerical	50
29	Final Software Verification Effort Hours	Numerical	50

30	Final Software Validation Effort Hours	Numerical	50
31	Final Software Review Effort Hours	Numerical	50
32	Final Software Audit Effort Hours	Numerical	50
33	Final Software Problem Resolution Effort Hours	Numerical	50
34	Final Cybersecurity Effort Hours	Numerical	50

In layman’s terms, there is no standard approach taught to defense cost estimators in what data values to use or not use in their data matrix when the missing, incomplete, or unreliable data values appear (DAU, 2018a). By offering the engineering managers and cost estimators within the discipline a standard approach to determine how to handle missing, incomplete, or unreliable data values, this could reduce the number of flawed cost estimates that lead to program cost overruns and unplanned additional federal budget request. Moreover DAU, the corporate university that was established to train and certify the Defense Acquisition Workforce (DAW) Business—Cost Estimating career field, does not train their students as to how to handle data sets when data is missing (DAU, 2018a). There is indeed a gap that needs to be addressed within the literature of the U.S. defense cost estimation body of knowledge that describes how defense cost estimators should

handle its unreliable and incomplete data problem when historical data sets have missing values (DAU, 2018a; GAO, 1972, 2009; Mislick & Nussbaum, 2015).

According to a U.S. defense based *Joint Agency Cost Estimating Relationship (CER) Handbook* (2018), it acknowledges that “data sets with missing and incomplete data” is a data analysis challenge and states that the “best course of action is to first attempt to remedy the problem by collecting more data, finding the information from the collected data set, and determining the cause of the unusual observations, respectively” (p. 221). This government document also acknowledges that it is “not always possible to correct such errors” and that it is important for estimators to understand the implications of these challenges, and to proceed with their analysis under caution (Joint CER Handbook, 2018, p. 221). The literature does not inform how cost estimators who leverage the defense Business—Cost Estimating curriculum at DAU directly handle the unreliable and incomplete engineering project data matrix problem other than through recognizing the problem through defense government documents and making changes to public policy (GAO, 2020).

The specific management problem is that there is a lack of research into the techniques to handle the unreliable and incomplete data problem. Consequently, U.S. defense cost estimators do not have an optimal set of options available to them when they must handle the unreliable and incomplete data problem when building a cost estimate, which allows them to forecast life-cycle cost analysis for a myriad of engineering-based acquisition projects and programs (DAU, 2018a; GAO, 2009, 2020; Morin, 2017). In a government publication memo entitled DoD Cost Analysis Data Improvement, Morin’s

approach to the problem was to start a data collection effort through updating eight topics withing two policies to improve data quality and estimation conditions (Morin, 2017).

This approach supports that research into cost analysis data quality is significant, and that improvements are still needed. There is a gap that needs to be addressed within the literature of the U.S. defense cost estimation body of knowledge that describes how defense cost estimators should handle its unreliable and incomplete data problem when historical data sets have missing values (DAU, 2018a; GAO, 1972, 2009; Mislick & Nussbaum, 2015).

How Other Disciplines Handle Unreliable/ Incomplete Data Problems

Through the lens of missing data theory, several empirical researchers have addressed the needs of both social and natural scientists across many disciplines with options to deal with handling their data matrices that may have unreliable, incomplete, or even completely missing values (Aittokallio, 2009; Baraldi & Enders, 2010; de Leeuw, 2001; García-Laencina et al., 2010, 2013; Tsirikitis, 2005). Moreover, many empirical researchers have assessed predictive accuracy on data matrices and have conducted experimental designs using missing data theory (Lin & Tsai, 2019). Unfortunately, none have been applied to any U.S. defense cost estimation data matrices and only a handful have used an experimental design to test missing data theory within in the software domain (Khoshgoftaar & Van Hulse, 2008; Song et al., 2008; Van Hulse & Khoshgoftaar, 2014). Currently, there is a gap that needs to be addressed within the literature of the U.S. defense cost estimation discipline that describes how defense cost estimators could handle its physical unreliable and incomplete data problem when

historical data sets have missing values (Brown & White, 2017; DAU, 2018a; GAO, 1972, 2009; Mislick & Nussbaum, 2015).

Could Missing Data Theory Improve Data Reliability/Completeness?

Addressing the problem of dealing with the problem of missing values in a representative DoD cost estimation when a physical data set has missing values has been ignored within the U.S. defense cost estimation discipline. A reliable and complete data matrix is a fundamental requirement to build a cost estimate, or even a cost estimation relationship (CER) model based on software effort hours and estimated software lines of code (ESLOC). When data values are not there, DoD cost estimators should have actionable techniques in which they could handle dealing with missing values vice relying on policies, statutes, and regulations of the environment to be the sole addressor of the specific problem (DAU, 2019a; DAU 2019b; Morin, 2017). Since cost estimation models' most important attribute is their forecasting accuracy, could applying missing data theory to missing values improve the disciplines' unreliable and incomplete data problem?

As a result of this literature review, the purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. At least two evaluative measures were used to test the impact of missing data theory techniques. Each

data set used in this study will have data values removed-at-random and at various percentages to create a simulation of the missing data problem applied to representative defense cost estimating data. The relationship between the removed-at-random-data-values from each data matrix group were operationally named the “Original Numerical Value” (pretest value) and the predicted value created as a result of applying a missing data theory technique were operationally named the “Predicted Numerical Value” (posttest value) for each data set’s experiments. The pretest and posttest value were compared by determining the average absolute error and relative error calculation to determine each missing data theory’s level of predictive accuracy.

Summary and Conclusions

A study to test if missing data theory techniques can solve the estimator’s unreliable and incomplete data problem was supported in the literature. In the first main section of this literature review, I provided the background surrounding the general data quality problem and U.S. public policy requirements that have been levied on the systems engineering economics subdiscipline of the Business—Cost Estimating discipline (GAO, 2009; Jorgenson, 2006, Morin, 2017). In the next section of my review of the literature, I narrowly focused and described the ineffectiveness of U.S. policy requirements that have been put in place to resolve the data reliability and completeness problem. This led me to the identification of the specific management problem that there is a lack of research into the techniques to handle the unreliable and incomplete data problem. Consequently, I described the U.S. defense cost estimators, engineering economists, and engineering managers problems they face in not having any options prescribed to them that can

address their data quality challenges beyond depending on policy enactments (Morin, 2017; 10 U.S. Code § 2334, 2017; 10 U.S. Code § 1746, 2012). I then assessed the literature and described common approaches to handling missing data that is used in other disciplines that have also faced a similar unreliable and incomplete data problem by using a theoretical approach. In the final main section, I described the gap in the literature and a need for U.S. defense cost estimators, engineering economists, and engineering managers to have options beyond policy to improve their data quality problem when they work with and assess physical data sets that are unreliable and incomplete for them to use for estimation. Lastly, I concluded the literature review's final section with why this capstone study to empirically test and measure the predictive accuracy of missing data theory techniques by applying it to the Business—Cost Estimating discipline is needed. This research can help determine which theoretically based technique(s) renders U.S. defense cost estimation data matrices and data sets most reliable and complete. Since non-defense software project and measurement data as it relates to cost estimation is written about in the literature, this study can extend knowledge in this area by applying empirical pre-experimental research to test and measure missing data theory techniques with representative U.S. defense cost estimation data.

Based on the gap discovered in this literature review, this study was necessary because “reliable and comprehensive cost data is essential to produce credible cost estimates as required in both (policy) statute and regulation” (Morin, 2017, p.1). In the past, GAO reported that the federal defense department lacked the data, both in volume and quality, needed to conduct effective cost estimates (Brown & White, 2017). As a

result, cost estimate realism to support future engineering systems' (e.g., developed software, aircraft, ships, business systems, autonomous systems, artificial intelligent systems etc.) success is threatened. This study provided a different perspective on an established problem that historical databases contain substantial amounts of missing data (Strike et al., 2001). By helping cost estimators, engineering economists, engineering managers, and even database administrators in the federal defense department "improve analyst productivity, quality of cost estimates, close data gaps, and provide the cost, acquisition, and resource allocation organizations with data required for better analysis and decision-making", an improvement to fund programs to an improved accurate estimated planned amount to complete an engineering project would be significant for these types of individuals (Morin, 2017, p. 1).

In conclusion, this capstone research sought to determine which missing data theory technique best lends itself to this inaugural body of research for the U.S. defense cost estimating discipline, with a high potential to influence future curriculum. I empirically determined what the predictive accuracy of traditional missing data theory techniques were when applied to a nonproprietary software measurement and engineering project data used in the U.S. defense cost estimating discipline. The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing

value percentages. The positive social change outcome of this research was that it helped determine that missing data theory techniques could provide options beyond public policy to address data quality problems and improve the state of the U.S. defense cost estimation discipline, discussed further in Chapter 3.

Chapter 3: Research Method

I structured this study for me to look at differences in the predictive accuracy measurements of three traditional missing data theory techniques within a pre-experimental design. Due to the specificity of the RQ, I used a purposive sampling of U.S. defense cost estimating representative data, in which the selected sample size of software programs was later randomly selected to receive an intervention. Found literature within engineering economics, software engineering economics, and cost estimation has not tested this theory or its applied techniques on U.S. defense cost estimation data. In addition, the U.S. defense cost estimation discipline had not addressed its current problem with any applied missing data theory techniques as an option to address its unreliable and incomplete data problem beyond changing its policies. The comprehensive purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. The positive social change outcome was to ultimately fill the gap in the literature regarding testing missing data theory's predictive accuracy applied to the U.S. defense cost estimator's unreliable and incomplete data problem. This study provided a quantitative analysis, and an understanding of the impact missing data theory could have in solving U.S. defense cost

estimator's current and longstanding problem that was recognized as early as 1972 (see GAO, 1972).

Chapter 3 includes the research design method, theoretical method of inquiry, justification of the research method, the justification of the intended sample and sample size, method of data collection and procedures, data management, data analysis technique and research method, issues of ethical considerations, reliability and validity, and instrumentation. As stated in Chapter 1, the RQ that grounds this study was: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? Being able to respond to this question from this research is important to provide the knowledge and understanding of missing data theory's potential to improve the data matrices used in cost estimation, especially in the U.S. defense industry's state of cost overruns and budget constraints.

Research Design and Rationale

The pretest value was operationally named the "Original Numerical Value" and represents the removed-at random-data-values that were obtained from the Walden University IRB approved purposive sample data that was used for empirical testing. The posttest value was operationally named the "Predicted Numerical Value" and represented the result of the treatment, the applied missing data theory technique, in which I was able to empirically evaluate and measure the predictive accuracy of the treatment's outcomes on a representative U.S. defense cost estimation data set by measuring its absolute error

and relative error values. The independent variables that were used to manipulate this quantitative pre-experimental research design of 30 out of 50 analogous and synthetic software development programs from the purposive data sample were the percentage of missing data (independent variable 1), the missing data theory technique (independent variable 2), and the 28 types of numerical data sets available within the one group data matrix. This study did not have a covariate, mediating, or moderating variable to account for as the subjects within the intervention are data values and not human beings.

The research design of this pre-experimental design was the best option to use for this empirical intervention study that tested theory and measured its results because it provided a controlled environment and provided pure construct validity to unequivocally answer the RQ effectively. In addition, like studies like this one that tested missing data theory, this research design provided the right amount of control to operationalize each independent and dependent variable, and yielded the data needed to evaluate the hypotheses, and answer the study's RQ (see Conte et al., 1986; Briand et al., 2000; Jeffery et al., 2000, 2001; MacDonell & Shepperd, 2003; Mittas & Angelis, 2008). Moreover, the pre-experiment introduced minimal time and resource constraint that allowed the ability to answer the RQ by conducting pre-experimental interventions at various missing data percentages, and allowed me to (a) test the predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, (b) compare them to complete data matrices used in the U.S. defense cost estimation discipline, and (c) determine which theories render incomplete and missing data matrices most reliable and complete. Each purposive data set from the one group

data matrix used in this study had data values removed-at-random to test the theory at various missing value percentages to create a simulation of the missing data problem applied to a representative U.S. defense cost estimating data.

There were two evaluative measures used to test the impact of missing data theory techniques by comparing pretest and posttest values. The first evaluative measure, the measure of predictive accuracy in the 28 treated data set were computed based on the delta change between the pretest and posttest values found before and after the treatment. This calculation determined the two dependent variables numerical values, the absolute error and relative error, that helped to compare the pretreatment and posttreatment data sets to measure each missing data theory's technique predictive accuracy, an aggregate needed to conduct ANOVA testing to determine if the study's results were significant. The second evaluative measure, significance testing, was performed by conducting a two-way ANOVA with an assumption of normality for a repeated measures ANOVA (see Field, 2018). The two-way ANOVA determined if there is an interaction between the missing data theory technique chosen, the multiple data set types from a representative U.S. defense cost estimation data matrix, and the various percentage levels of missing data (i.e., this pre-experimental study's three independent categorical variables) to explain the measures of predictive accuracy (i.e., the two dependent/outcome numerical variables) using absolute error and relative error calculations once all experiments have been completed (see Field, 2018). Furthermore, this significance testing used a two-way (within-subject variables) and three between-subjects factors analysis of variance

(ANOVA) to measure the effects and interactions that existed between the independent variables and dependent variables (see Field, 2018).

The pre-experimental design, chosen because of the need to experiment on a purposive sample of a representative U.S. defense cost estimation data, leveraged a complete as well as logical data matrix, and measured the accuracy of applied missing data theory techniques results after the artificial missing data problem was created (Trochim & Donnelly, 2008). Artificially inducing the missing values into the data matrix had been chosen because it is an approach that had been adopted in several missing data experimental studies, and allowed me the aptitude to apply several missing data theory technique treatments to a sample, and test for predictive accuracy of this intervention while adding knowledge and understanding to the gap in the literature (see Brown & White, 2017; DAU, 2018a; GAO, 1972, 2009; Hill, 2011; Mislick & Nussbaum, 2015). By empirically evaluating missing data theory techniques' predictive accuracy, a new method to handle missing data problems for DoD cost estimators, engineering managers/economists, database administrators, data scientists, as well as other researchers could be a welcomed addition to the Defense Acquisition University's (DAU's) Business—Cost Estimating current curriculum, as well as adds to the engineering economics scholarly conversation.

Methodology

I have just described the research design method and have defined the theoretical method of inquiry and justification for this research method. In the next section, I describe the intended population, as well as the justification for the intended sample. In

addition, I justify the sample size based on the unique population of interest. In addition, I provide the sampling procedures that I conducted for this intervention study. Lastly, the details about the use of archival data, and procedures to take once the data collection occurs will be described in more detail.

Population

Each year, the DoD captures its list of Major Automated Information Systems, all high visibility and high dollar information system programs. As of October 1, 2018, the Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Acquisition Resources and Analysis (ARA) Directorate published its annual 2019 MDAP and Major Automated Information Systems (MAIS) list. To properly focus this study on software effort-based cost estimation, the population in which the pre-experimental design is in support of is the current 30 DoD software systems identified, and any future MAIS systems identified in later years (Undersecretary of Defense for Acquisition and Sustainment, 2019). Table 2 captures the targeted population in which this quantitative study is in support of. The Category (Cat) describes if the acquisition program has oversight at the Component Level (IAC) or at the DoD Level (IAM) based on its budgetary significance and risk (Undersecretary of Defense for Acquisition and Sustainment, 2019).

Table 3

DoD Major Automated Information Systems (MAIS) List

Id	Short Name	Long Name	Component (Cat)
1	ACWS	Army Contract Writing System	Army (IAC)

2	AFIPPS Inc 1	Air Force Integrated Personnel and Pay System Increment 1	Air Force (IAC)
3	BEC Inc 1	Biometrics Enabling Capability Increment 1	Army (IAC)
4	CAC2S Inc 1	Common Aviation Command and Control System Increment 1	Navy (IAC)
5	CANES	Consolidated Afloat Networks and Enterprise Services	Navy (IAC)
6	DAI Inc 2	Defense Agencies Initiatives Increment 2	DLA (IAM)
7	DAI Inc 3	Defense Agencies Initiatives Increment 3	DLA (IAM)
8	DCAPES Inc 2B	Deliberate and Crisis Action Planning and Execution Segments Increment 2B	Air Force (IAM)
9	DCGS-A Inc 1	Distributed Common Ground System -Army Increment 1	Army (IAC)
10	DCGS-A Inc 2	Distributed Common Ground System -Army Increment 2	Army (IAC)
11	DCGS-N Inc 2	Distributed Common Ground System -Navy Increment 2	Navy (IAC)
12	DEAMS Inc 1	Defense Enterprise Accounting and Management System - Increment 2	Air Force (IAM)
13	DEOS	Defense Enterprise Office System	DISA (IAM)
14	DHMSM	Department of Defense Healthcare Management System Modernization	DHA (IAM)
15	ESBMC2	Enterprise Space Battle Management Command and Control	Air Force (IAM)
16	GCSS-A Inc 1	Global Combat Support System -Army Increment 1	Army (IAC)

17	GCSS-A Inc 2	Global Combat Support System -Army Increment 2	Army (IAC)
18	GCSS-J Inc 8	Global Combat Support System -Joint Increment 8	DISA (IAC)
19	IPPS-A Inc 2	Integrated and Personal Pay System -Army Increment 2	Army (IAC)
20	ISPAN Inc 4	Integrated Strategic Planning and Analysis Network Increment 4	Air Force (IAC)
21	ISPAN Inc 5	Integrated Strategic Planning and Analysis Network Increment 5	Air Force (IAC)
22	JMS Inc 2	Joint Space Operations Center (JSpOC) Nissin System Increment 2	Air Force (IAM)
23	JOMIS	Joint Operational Medicine Information Systems	DHA (IAM)
24	KMI Inc 2	Key Management Infrastructure Increment 2	NSA/CSS (IAC)
25	KMI Inc 3	Key Management Infrastructure Increment 2	NSA/CSS (IAM)
26	MPS Inc 5	Mission Planning System Increment 5	Air Force (IAC)
27	MROi	Maintenance Repair and Overhaul Initiative	Air Force (IAC)
28	PKI Incr II	Public Key Infrastructure Increment 2	NSA/CSS (IAC)
29	Teleport Gen 3	Teleport Generation 3	DISA (IAC)
30	TMC	Tactical Mission Command	Army (IAC)

Sampling and Sampling Procedures

Sampling is the process of selecting a representative group from a population to be studied. With the population of 30 major DoD MAIS programs' cost estimating data being contractor and acquisition sensitive, as well as requires signed non-disclosure

agreements (NDAs) from anyone who sees this data, I used nonproprietary synthetic data for this study that was and is a representative sample of what U.S. defense cost estimators work with to estimate cost (Undersecretary of Defense for Acquisition and Sustainment, 2019). With IRB approval, I used a purposive sample taken from a DAU course, BCF 250 Software Cost Estimation, that extracts data for its classroom exercises from the FACADE database repository (Trochim & Donnelly, 2008). In being a synthetic nonproprietary data matrix that is currently used to teach U.S. defense cost estimators, this sample was ideal for research that can be shared within the academic and scholarly literature. DAU data is within the public domain and captured under the IRB form category as data that is found within “public records or documents” (17 U.S. Code § 105, 2010). By using a data matrix used by professors to teach students, as well as one that is representative of the type of data in which many U.S. defense cost estimators are exposed to while seeking certification in the discipline, this purposive sample of nonproprietary, synthetic, and software effort estimation synthetic data from the FACADE database repository was used. Since purposive sampling, a nonprobability sampling procedure, was used in this study to test missing data theory techniques on a niche population, a power analysis to determine sample size and effect size was not prudent (see Trochim & Donnelly, 2008; UCSF, 2019). The sample size for this pre-experimental research was comprised of 30 out of 50 analogous and synthetic software development programs and was able to inform the population of 30 DoD MAIS programs shown in Table 3. Stated differently, one group of 30 software development programs, comprised of 28 numerical data sets in a data matrix was used to simulate the U.S. defense cost estimator’s missing

data problem and to make inferences about how this population may respond to missing data theory treatments.

Procedures for Data Collection (Purposive Sample of Archival Data)

In narrowing this research study to have positive social change impact, I contacted the director of academic program at DAU. This university is the corporate university for U.S. DoD cost estimators as established under 10 U.S. Code § 1746 (2012). This law states that “The Secretary of Defense... shall establish and maintain a defense acquisition university structure to provide for the professional educational development and training of the acquisition workforce” (10 U.S. Code § 1746, 2012, ¶. 1). With this university serving as the Defense Acquisition Workforce Improvement Act Level I, II, and III Acquisition Professional and Development Program certifying agent for the Business—Cost Estimating discipline, experimenting on its curriculum’s archival datasets with an intervention of applied statistical missing data theory techniques is ideal for this pre-experimental pretest-posttest design. The archival data, also termed synthetic data, represents U.S. defense cost estimation data that is not contract or acquisition sensitive, is used in curriculum to educate and train U.S. defense cost estimators in courses, and taken from the FACADE database repository (DAU, 2018b).

Intervention (One Group Pretest-Posttest Design/Pre-experimental)

For this study, the intervention was on a synthetic data, not a human, by providing three missing data theory techniques on a U.S. defense cost estimation representative data matrix to test the predictive accuracy of each data set found within the data matrix empirically. The one group pretest-posttest no control group/pre-experimental design

included 4,704 treatments in which I measured the outcome of each “Original Numerical Value” and calculated the delta values as a result of the missing data theory technique’s treatment resulting “Predicted Numerical Value” to measure the treatment’s level of predictive accuracy. Since the sample of software development programs within each of the 28 data sets did not fall below the DoD MAIS Program population size of $N=30$, this study’s use of a purposive sample of $n=30$ fictitious yet representative U.S. defense software development programs is robust.

I tested and measured missing data theory techniques level of predictive accuracy by empirically applying three techniques as an intervention on 28 artificially missing data sets comprised of 30 fictitious software development programs. I used traditional statistical approaches of missing data theory to establish the foundation of its utility to the Business—Cost Estimating discipline. In this study, I conducted an intervention on a synthetic missing value data matrix problem applied to a representative U.S. defense cost estimation data. I applied a post positivist worldview in which I tested and assessed three missing data theory techniques based on a quantitative pre-experimental design. The research used concepts from the traditional survey, one group pretest-posttest no control group/pre-experimental research design by taking the following approach (see Thyer, 2012):

1. Take a synthetic complete data set made up of at least 30 software development programs and 28 numerical (quantitative) software data set types (independent variable 2) in which, for example, are the software effort hours

that have been collected and represent a U.S. defense cost estimation data matrix.

2. Document the “Original Numerical Value”, the original data set, that represents the software effort attribute that has been collected in a database as a pretest measurement point for all 30 programs in the complete data set.
3. Calculate univariate descriptive statistics, to include the mean, for each data set in the data matrix (e.g., to describe Data Set Type 14 Final Software Requirements Analysis Effort Hours for the 30 software development programs that were used in this study).
4. Assuming the mechanism of MCAR, simulate the missing data problem by beginning the random removal of programs to create an arbitrary missing value pattern within each complete data set at 8 percentage levels of missingness. Begin by randomly removing only 5% of the data values by utilizing a random number generator to remove values based on where the data value was originally placed within the complete data set (Research Randomizer, 2020).
5. Examine the pattern of the artificially induced missing data set, document the data set type (independent variable 2) and document the missing data percentage (independent variable 1).
6. Apply the respective missing data theory techniques (independent variable 3) to the artificially induced data set and impute values that were missing utilizing the IBM SPSS 25 statistical package. The application of the

following missing data theory technique will serve as the intervention to test this study's null and alternative hypotheses with applying the following treatments:

- Complete Case Analysis (Toleration)/ Listwise Delete Methods;
 - Single Imputation Methods (Direct) Methods (García-Laencina et al., 2010);
 - Multiple Imputation (Direct) Methods (Enders, 2006).
7. Calculate the absolute error and relative error to compare the pretest value ("Original Numerical Value" = A) and posttest value ("Predicted Numerical Value" = B) results after the application of the missing data theory technique treatment that was artificial simulated to have 5% of the data set to contain missing values.
 8. Calculate univariate descriptive statistics based on the post-treatment data set that has been made 100% complete with new "Predicted Numerical Values" as a result of the applied missing data theory technique that predicted and replaced 5% of the numerical values that were artificially missing from the original data set.
 9. Compare calculate univariate descriptive statistics between the pre- and the post- treated data set that has been made complete from the application of missing data theory to measure the predictive accuracy of missing data theory technique applied.

10. Repeat steps 2-9 for each of the missing data value percentages seven more times (e.g., 10%, 15%, 20%, 25%, 30%, 35%, 40%) to test predictive accuracy at various levels as tested in the literature for 28 data sets from a representative U.S. defense cost estimation data matrix.

Archival Data

The archival data set used for this one group pretest-posttest no control group/pre-experimental design to test the RQ was an IRB approved data matrix made available via a webpage from the DAU's BCF 250 Software Cost Estimation course. The data matrix used was an excerpt from an in-classroom exercise to analyze the FACADE historical database for completed software efforts and is in the public domain. With a significant amount of the scholarly literature looking at software effort estimation data, testing the RQ and hypotheses on a U.S. defense software effort cost estimation data matrix allowed this body of work to enter the larger scholarly conversation on the topic. The targeted archival complete data matrix was comprised of 30 out of 50 analogous and synthetic software development programs with 28 data sets related to software effort, size and schedule data used in software cost estimation. Each of the 30 synthetic software programs have observations on 15 software effort attributes, 12 software size attributes, and one external software interface requirement attributes that make up 28 data sets out of the single data matrix.

Instrumentation and Operationalization of Constructs (IBM SPSS 25)

The instrumentation used to answer the RQ by testing and measuring the associated hypotheses was IBM SPSS 25 licensed software. Due to the statistical

functionality of IBM SPSS 25 software, its instrumentation was used to administer missing data theory techniques as a treatment and measure its effects on U.S. defense cost estimating data within a pre-experimental design. Missing data theory techniques require computational power that is provided in statistical and machine learning software tools, in which IBM SPSS 25 includes. As a foundational study for U.S. defense cost estimators, testing traditional missing data theory techniques that is supported within the Missing Value Analysis module and Multiple Imputation functionality within IBM SPSS 25 were used. SPSS is recognized in the academic community as reliable and valid and has the processing power to assess data that has incomplete and missing values (Enders, 2010). The Missing Value Analysis module and Multiple Imputation functionality in IBM SPSS 25 has the computational and algorithmic ability to compute traditional missing data theory techniques that can be applied to the first study that applies these techniques to the U.S. defense cost estimation domain. This study tested and measured the level of predictive accuracy of traditional missing data theory techniques that include a) Complete Case Analysis or Listwise Deletion (LD), b) Single Imputation using the Mean (SI-Mean), and c) Multiple Imputation using Linear Regression (MI-LR) techniques on an IRB approved and representative U.S. defense cost estimation data matrix.

Intervention Studies or Those Involving Manipulation of Independent Variables

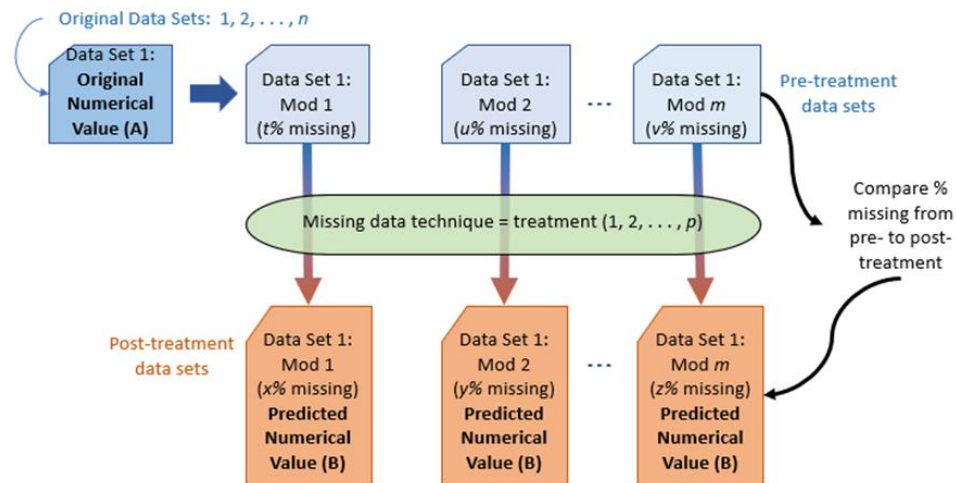
For this intervention on the pretest value, the “Original Numerical Value”, the following parameters were manipulated to simulate the U.S. defense cost estimation missing data problem (Aljuaid & Sasi, 2016; Strike et al., 2001):

- The data set type from a data matrix (independent variable 1).
- The percentage of missingness of the observations/programs with missing data (independent variable 2), based on the assumption that any missing observations are MCAR.
- The missing data theory technique (independent variable 3).

Eight different percentages of missing values were simulated per each of the missing data theory techniques applied (e.g., 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%). It is generally accepted that data sets with more than 40% missing data are not useful for detailed analysis (Strike et al., 2001). Twelve experimental studies considered missing data percentages above 50% in which they all had very large data sets of several hundreds to 12,000 (Chen et al., 2017; Eirola et al., 2013; Graham et al., 2007; Janssen et al., 2010; Kapelner & Bleich, 2015; Kiasari et al., 2017; Li et al., 2014; Li & Parker, 2014; Mesquite et al., 2017; Purwar & Singh, 2015; Qin et al., 2009; Zhu et al., 2010). Since U.S. defense cost estimation data levels of analogous systems are not large data sets, this study will not exceed a missing data percentage rate beyond 40%. Figure 2 is a full schematic of this study in which it defines the independent and dependent variable definitions, and as well as depicts how pretest and posttest approximation errors were calculated.

Figure 2

Statistical Method to Perform this One Group Pretest-Posttest Design



IVs (3):

Data set (categorical): 1, 2, ..., n
 Mod (categorical): various % of missing data, t, u, ... v%
 Missing data technique (categorical): 1, 2, ..., p

DVs (2):

Measure of predictive accuracy in treated data set (numerical)
 Absolute error (interval) = $(A - B)$ [e.g., $90 - 75 = 15$]
 Relative error (interval) = $(A - B) \div B$ [e.g., $(90 - 75) \div 75 = 20\%$]

Data Analysis Plan

Data was analyzed from the results of 4.704 pre-experimental treatments in this one group pretest-posttest no control group/pre-experimental design. All numerical values generated from the pretest and posttest numerical values used the ratio/scale of measurement. I used absolute error and relative error calculations, approximation error terms, to measure the predictive accuracy of each application of traditional missing data theory techniques (Kreinovich, 2012). Seo et al. (2009) used Magnitude of Relative Error (MRE) and Magnitude of Error Relative (MER) calculation of each to measure the software effort estimation predictive accuracy in their study as follows (p. 3):

$$MRE_i = \frac{|ActualEffort_i - EstimatedEffort_i|}{ActualEffort_i},$$

$$MER_i = \frac{|ActualEffort_i - EstimatedEffort_i|}{EstimatedEffort_i}$$

These calculations are consistent with the logic needed to measure missing data predictive accuracy, using approximation error as the type of calculation to use. Likewise, many other researchers have used approximation error as an evaluation criteria to measure the predictive accuracy of their effort estimation models as well (see Conte et al., 1986; Briand et al., 2006, 2000; Jeffery et al., 2000, 2001; MacDonell & Shepperd, 2003; Mittas & Angelis, 2008).

To answer the RQ and test and measure this study's hypotheses, using approximation error-based equations, I measured the delta of the single group's pretest and posttest values based on the intervention of each missing data theory technique treatment, which represents the most critical step in this pre-experimental study's analysis. The absolute error and relative error calculations captured the outcome variables needed to assess the findings from 4,704 treatments and served as the study's two dependent variables (absolute error and relative error) to measure the level of predictive accuracy as it interacts with the independent/predictor variables that have been manipulated in order to answer the RQ. This data analysis plan will support answering the RQ: To what degree can traditional missing data techniques accurately solve cost estimators' and engineering manager's unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix?

IBM SPSS 25 and Microsoft Excel were the software used to analyze and evaluate the results of this one group pretest-posttest/pre-experimental design in a systematic and consistent manner to calculate the level of predictive accuracy per each data set type at each percentage level of missingness. Significance testing was then performed by conducting a two-way repeated measures ANOVA. A repeated measures ANOVA “is a term used when the same cases participate in all conditions of an experiment” in which at least two or more variables manipulate the experiment (Fields, 2018, p. 651). The *F* ratio from this analysis will explain the main effects and interaction between all the independent variables and dependent variables. Covariate variables were not used considering the single group of software program characteristics was data and not human beings who may have outside variables to impact experimental results. The data sets from the data matrix used for all the pre-experiments were kept constant. This quantitative research method of inquiry has been chosen to help determine how well defense cost estimators could handle historical data sets with the use of missing data theory techniques (Kirk, 2013).

Furthermore, the RQ for this study was to determine what the predictive accuracy is from various missing data theory techniques when applied to defense cost estimating data matrices. The RQ is as follows: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? The null and alternative hypotheses that was used to answer the RQ was derived from the results of the sole data matrix using a one group pretest-

posttest no control group/pre-experimental design. The calculated measure of predictive accuracy (e.g., error approximation) provided a table of before and after average absolute and relative error values because of the applied three treatments of missing data theory techniques. There were 28 data sets each comprised of 30 analogous and synthetic software development programs as the foundation for this pre-experiment predictive accuracy evaluation. After which, an ANOVA was conducted to explain the interaction of this study's variables using the following null and alternate hypotheses:

H_01 : There are no significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, the Complete Case Analysis/ Listwise Delete approach?

H_a1 : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, Complete Case Analysis/ Listwise Delete approach? The means are not equal.

H_02 : There are no significant differences evident between the data set's mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 2, a Single Imputation approach?

H_a2 : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in

comparison to those that are computed “Predicted Numerical Values” using missing data theory 2, a Single Imputation approach? The means are not equal.

H₀₃: There are no significant differences evident in the data sets’ mean absolute error and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 3, the Multiple Imputation approach?

H_{a3}: There are significant differences evident between the data sets’ mean absolute error and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 3, the Multiple Imputation approach? The means are not equal.

Threats to Validity

The research design of this pre-experimental design was carefully chosen to address external validity, internal validity, construct validity, and ethical challenges. I pursued an experimental design since it provided the structure needed to plan and manage my approach to academic inquiry. Since an experimental controlled environment could unequivocally help me to answer the RQ effectively, I assessed what could be done to ensure the integrity of my study was sound. The next sections will describe how I leveraged literature and designed an experiment to mitigate threats to this study’s validity.

External Validity

In order to ensure that the most relevant primary studies were being included keywords related to missing values as well as terms related to experimental designs and software cost estimation were used in the search string to discover a wide range of papers covering empirical studies in respect to applied missing data theory. However, some terms may have been missed in the search string which could have affected the results of this paper, and the study undertaken. This issue would only have a minor influence since I used different libraries and scanned references of relevant papers in order to minimize the risk of missing any relevant materials to exhaust the literature search.

In order to present relevant results that can be exploited by other researchers, the search string, the databases and the inclusion/exclusion criteria and every step performed to focus the research was presented in the Literature Search Strategy in Chapter 2.

Internal Validity

Internal validity was established by selecting instrumentation that was leveraged by many practitioners in the social sciences for statistical purposes. Even though using it as instrumentation to test and measure the techniques of missing data theory, its longevity and annual updates on both the IBM Missing Value Analysis and Missing Data Analysis modules in IBM SPSS 25 are up to date and meet the needs of internal validity.

Common threats to pretest-posttest designs typically come from what has been found during social science experiments that consist of human being subjects. Typically, threats such as history, statistical regression, subject fatigue all distort the results and the internal validity in a study of this kind (Shek & Zhu, 2018). In the case of having data

serve as the subject, internal validity is less threatened. In general, the pretest–posttest design is useful in intervention studies and program evaluations when it is well conducted and when the researcher is cautious in drawing causal inferences from its results. There are two common ways to strengthen the pretest-posttest design. First, if all measures consistently change in a predicted direction after an intervention, using several instead of just one valid and reliable outcome measure can make conclusions more convincing about a study. Secondly, multiple pretests and multiple posttests can provide more credible evidence regarding the participating human being or thing (e.g., software programs' estimation attribute) before and after an intervention (e.g., the treatment of missing data theory techniques) to inform results that are both immediate and long-term outcomes. In fact, if a series of pretests and posttests are employed over a longer timeframe, a one group pretest–posttest/ pre-experimental design would change into a quasi-experimental scheme known as the interrupted time series (ITS) design which is considered a stronger study (Thyer, 2012). My execution of this pre-experimental design looks at multiple pretests and posttests outcome variables and repeats the intervention treatment 4,704 times as a means to strengthen this study's internal validity beyond the instrumentation used.

Construct Validity

This pre-experiment measured what it purported based on leveraging approaches in the literature to evaluate the level of predictive accuracy of missing data theory technique treatment results on a representative U.S. defense cost estimation data matrix. Specific to this intervention study, missing data theory techniques (independent variable

3), representative data sets (independent variable 1), and artificially inducing the missing data problem at eight different percentages (independent variable 2) of missing data to evaluate the RQ: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering manager's unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? Constructing the research to ultimately measure predictive accuracy by controlling the research design after acquiring a representative U.S. defense cost estimation complete data matrix was a requirement. In addition, using a random number generator to remove values, applying and measuring each missing data theory technique's treatment and posttest value, and running statistical tests to fully evaluate the results will fill the gap in the current literature of the U.S. defense cost estimation discipline. The measure of predictive accuracy and leveraging approximation error was used to evaluate accurate software effort estimation in other studies and was used to directly answer the RQ for this study (see Conte et al., 1986; Briand et al., 2000; Jeffery et al., 2000, 2001; MacDonell & Shepperd, 2003; Mittas & Angelis, 2008).

Ethical Procedures

The research study's participants were limited to a U.S. defense cost estimation data matrix and has zero impact on human subjects. The research was ethical and socially sound since it used data that was nonproprietary to any defense contractor and does not violate any acquisition sensitive laws. The archival data matrix is held by DAU, per Title 17 U.S. Code § 105 (2010) and is in the public domain for use. The pre-experimental one group pretest-posttest design procedures are well within ethical standards to test the RQ

that states: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering manager's unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix? The proposal for this dissertation and appropriate documentation was submitted and approved by the IRB to proceed to final study (Walden University, 2020).

Summary

Missing data theory provides promising techniques that could be incorporated into the defense cost estimation discipline for practitioners. A one group pretest-posttest no control group/pre-experimental research design using a representative U.S. defense cost data matrix was exposed to intervention techniques that were grounded in missing data theory. Experimental as well as pre-experimental design is the backbone of good research and was found within the literature as an approach to continue this academic conversation with other scholars who are engaged in empirical software engineering and cost estimation research and findings. Pre-experimental designs do not require a controlled environment but does require an isolation of variables. As a result, this type of experimental design is as an applicable design for the analysis of applied missing data theory.

Chapter 4 contains the results after 4,704 treatments were conducted using this type of experimental research design that compared pretest and posttest numerical values that detail the level of predictive accuracy of applied missing data theory techniques. This study's findings as a measurement of predictive accuracy can later become a quasi-experimental if an interrupted time series/longitudinal study is continued for this exact

study (Campbell & Stanley, 1963; Cook & Campbell, 1979; Reichardt, 2019; Shadish et al., 2002; Thyer, 2012). This pre-experimental study answers the RQ and tests the hypotheses that helps determine if missing data theory is a viable approach to handling the missing data problem within the U.S. defense cost estimation Business—Cost Estimating discipline.

Chapter 4: Results

The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. The three independent variables used for this study were the different percentage levels of missingness created (independent variable 1), the category title of the data set type (independent variable 2), and the traditional missing data theory techniques (independent variable 3). The two dependent variables used for this study were the absolute errors and relative errors calculated from the pre-experimental treatments derived from the data sets' pretest and posttest numerical values. Differences in the absolute error and relative error groups were assessed by a two-way repeated measures ANOVA testing on the pretest and posttest values. I used eight different percentages for missing values (diminished completeness) with three missing data theory techniques. For the individual treatments, I randomly selected a subset from my purposive sample down to 30 software programs that were analogous to each other. Stated differently, these 30 software programs were similar to each other in that they all were (a) air vehicle software applications for a sensor control and signal processing operational environment, (b) were a part of an air vehicle system, (c) leveraged the waterfall software development paradigm, and (d) were all software upgrades to current software on its host air vehicle. Each selected software program had 28 numerical data

sets either related to the software effort, size, and number of software interfaces is used in the Business—Cost Estimating discipline to infer what a future analogous software development engineering project may cost.

The RQ for this study was designed to investigate what the predictive accuracy was from various missing data theory techniques when applied to a defense cost estimating data matrix. The null and alternative hypotheses that was used to answer the RQ was derived from the results of the sole data matrix using the one group pretest-posttest no control group/pre-experimental design. The calculated measure of predictive accuracy (e.g., error approximation value) provided a table of before and after average absolute and average relative error values because of the applied three treatments of missing data theory techniques. There were 28 data sets tested for each of the 30 out of 50 analogous and synthetic software programs found within the data matrix. After which, a two-way repeated measures ANOVA was conducted to explain the interaction of this study's two dependent variables using the following null and alternate hypotheses:

RQ: To what degree can traditional missing data theory techniques accurately solve cost estimators' and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix?

H_0 1: There are no significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, the Complete Case Analysis/ Listwise Delete approach?

H_{a1} : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 1, Complete Case Analysis/ Listwise Delete approach? The means are not equal.

H_{02} : There are no significant differences evident between the data set's mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 2, a Single Imputation approach?

H_{a2} : There are significant differences evident between the data sets' mean absolute and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 2, a Single Imputation approach? The means are not equal.

H_{03} : There are no significant differences evident in the data sets' mean absolute error and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 3, the Multiple Imputation approach?

H_{a3} : There are significant differences evident between the data sets' mean absolute error and mean relative error of actual values "Original Numerical Values" in comparison to those that are computed "Predicted Numerical Values" using missing data theory 3, the Multiple Imputation approach? The means are not equal.

Chapter 4 contains the details about the data collection effort, treatment fidelity, results, and summary tables from the 4,704 treatments from the empirical research methodology implemented to answer this study's RQ. This chapter describes the execution of this empirical research through the one group pretest-posttest no control group/pre-experimental design used to investigate a representative U.S. defense cost estimation data matrix in the public domain.

Data Collection

As planned, I was able to obtain a representative U.S. defense cost estimating data matrix from the public domain made available through a DAU registered account and login once it was approved by the IRB at Walden University for this empirical research. In addition, I was able to gain immediate access to the BCF 250 Software Cost Estimation course materials and the Paired SRDR Database data matrix, in which contained software effort estimation data on 50 synthetic DoD MAIS programs. At that point, I simulated the cost estimator's variant of the missing data problem and made the data matrix representative of what was needed to develop a U.S. defense cost estimate. As a result, I down selected the number of 50 software programs to 30 software programs to have a single set of programs that were similar to each other. After analyzing the data, I was able to determine that 30 software programs shared the same operational environment, development paradigm, and software development phase. Table 4 displays which of the 30 programs were selected and display what the specific same characteristics were that the single group of analogous software programs had in common to be used in this quantitative study. The 30 out of 50 analogous and synthetic software

development programs became the single group for the one group pretest-posttest no control group/pre-experimental design.

Table 4

Selected Analogous Programs from DAU Data Matrix Based on Application Type (e.g., Operational Environment, Development Paradigm and Phase)

Program No. in Data Matrix Sequence	Assigned New ID for Experiment Order	Same Software Application Type (Sensor Control & Signal Processing)	Same Operating Environment (Air Vehicle)	Same Development Paradigm (Waterfall)	Same Development Phase (Upgrade to Current Software)
1	N/A	Yes	No	No	No
2	N/A	Yes	No	No	No
3	P1	Yes	Yes	Yes	Yes
4	N/A	Yes	No	No	No
5	P2	Yes	Yes	Yes	Yes
6	P3	Yes	Yes	Yes	Yes
7	N/A	No	No	No	Yes
8	N/A	No	No	Yes	No
9	P4	Yes	Yes	Yes	Yes
10	P5	Yes	Yes	Yes	Yes
11	P6	Yes	Yes	Yes	Yes
12	N/A	Yes	Yes	No	Yes
13	N/A	Yes	Yes	No	No
14	P7	Yes	Yes	Yes	Yes
15	N/A	No	Yes	No	No

16	N/A	Yes	No	Yes	No
17	P8	Yes	Yes	Yes	Yes
18	P9	Yes	Yes	Yes	Yes
19	N/A	No	No	No	Yes
20	N/A	Yes	Yes	Yes	No
21	P10	Yes	Yes	Yes	Yes
22	P11	Yes	Yes	Yes	Yes
23	P12	Yes	Yes	Yes	Yes
24	N/A	Yes	No	No	No
25	P13	Yes	Yes	Yes	Yes
26	P14	Yes	Yes	Yes	Yes
27	P15	Yes	Yes	Yes	Yes
28	P16	Yes	Yes	Yes	Yes
29	P17	Yes	Yes	Yes	Yes
30	P18	Yes	Yes	Yes	Yes
31	P19	Yes	Yes	Yes	Yes
32	P20	Yes	Yes	Yes	Yes
33	P21	Yes	Yes	Yes	Yes
34	P22	Yes	Yes	Yes	Yes
35	P23	Yes	Yes	Yes	Yes
36	P24	Yes	Yes	Yes	Yes
37	P25	Yes	Yes	Yes	Yes
38	N/A	No	Yes	No	Yes
39	N/A	Yes	Yes	No	Yes
40	N/A	Yes	Yes	No	No
41	N/A	Yes	Yes	Yes	No

42	P26	Yes	Yes	Yes	Yes
43	P27	Yes	Yes	Yes	Yes
44	P28	Yes	Yes	Yes	Yes
45	N/A	Yes	Yes	Yes	No
46	P29	Yes	Yes	Yes	Yes
47	P30	Yes	Yes	Yes	Yes
48	N/A	Yes	Yes	No	Yes
49	N/A	Yes	No	Yes	Yes
50	N/A	Yes	No	Yes	Yes

Similarly, the data collected from the data matrix also had a total of 34 data sets in which provided characteristics about the 30 synthetic DoD MAIS programs that were used for this study. Only 28 of the 34 data sets were selected for this study due to them being numerical data that could be removed-at-random to treat with various missing data theory techniques. Table 5 shows which of the 28 data sets were selected due to being numerical values for this one-group pretest-posttest no control group/pre-experimental design. The 30 out of 50 synthetic and analogous software develop programs and their corresponding 28 data sets used were representative of the population, the annual DoD MAIS list of all high visibility and high dollar information system programs. Not only did this study's sample size represent the population, but the data set came from course material used to certify cost estimators on ways to estimate software programs. This sample appropriately supports this intervention study in respect to software effort-based cost estimation, and the population in which the pre-experimental design is in support of. Based on the results and summary of the pre-experiments, current DoD software systems,

and any future DoD MAIS systems identified in later years could have more reliable and complete data available for cost estimators to use while estimating software program costs with better predictive accuracy.

Table 5

Selected Numerical Data Sets from DAU Data Matrix

Data Set in Matrix Sequence	New Id for Pre-Treatment	Data Set Type	Nominal or Numerical?	Select-ed for Experiment?
1	N/A	Software Intensive Program	Nominal	No
2	N/A	Mapped Application Type	Nominal	No
3	N/A	Operating Environment	Nominal	No
4	N/A	Primary Programming Language	Nominal	No
5	N/A	Development Paradigm	Nominal	No
6	N/A	Upgrade/New	Nominal	No
7	DataSet1	Number of External Interface Requirements	Numerical	Yes
8	DataSet2	Initial SLOC – New	Numerical	Yes
9	DataSet3	Initial SLOC – Modified	Numerical	Yes
10	DataSet4	Initial SLOC – Reused	Numerical	Yes
11	DataSet5	Final SLOC – New	Numerical	Yes
12	DataSet6	Final SLOC – Modified	Numerical	Yes
13	DataSet7	Final SLOC – Reused	Numerical	Yes

14	DataSet8	DM % - Modified*	Numerical	Yes
15	DataSet9	CM % - Modified*	Numerical	Yes
16	DataSet10	IM % - Modified*	Numerical	Yes
17	DataSet11	DM % - Reused*	Numerical	Yes
18	DataSet12	CM % - Reused*	Numerical	Yes
19	DataSet13	IM % - Reused*	Numerical	Yes
20	DataSet14	Final Software Requirements Analysis Effort Hours	Numerical	Yes
21	DataSet15	Final Software Architectural Design Effort Hours	Numerical	Yes
22	DataSet16	Final Software Detailed Design Effort Hours	Numerical	Yes
23	DataSet17	Final Software Construction Effort Hours	Numerical	Yes
24	DataSet18	Final Software Integration Effort Hours	Numerical	Yes
25	DataSet19	Final Software Qualification Testing Effort Hours	Numerical	Yes
26	DataSet20	Final Software Documentation Management Effort Hours	Numerical	Yes
27	DataSet21	Final Software Configuration Management Effort Hours	Numerical	Yes
28	DataSet22	Final Software Quality Assurance Effort Hours	Numerical	Yes
29	DataSet23	Final Software Verification Effort Hours	Numerical	Yes
30	DataSet24	Final Software Validation Effort Hours	Numerical	Yes

31	DataSet25	Final Software Review Effort Hours	Numerical	Yes
32	DataSet26	Final Software Audit Effort Hours	Numerical	Yes
33	DataSet27	Final Software Problem Resolution Effort Hours	Numerical	Yes
34	DataSet28	Final Cybersecurity Effort Hours	Numerical	Yes

Since purposive sampling, a nonprobability sampling procedure, was used in this study to test missing data theory techniques on a niche population, a power analysis to determine sample size and effect size was not prudent (Trochim & Donnelly, 2008). The sample size for this pre-experimental research was comprised of 30 out of 50 analogous and synthetic software development programs and was able to inform the population of 30 DoD MAIS programs of a generalized solution based on the testing and measurement of predictive accuracy. From an external validity perspective, randomly selecting the programs treated with three missing data theory techniques across the purposive sample added a probabilistic element to this pre-experimental design and strengthened the ability to make a stronger inference as to what missing data theory techniques offer the Business—Cost Estimating discipline the highest level of predictive accuracy.

Treatment and Intervention Fidelity

The intervention to apply missing data theory techniques to data sets from a data matrix went as planned. Some additional concepts were noted that do serve as a requirement to use missing data theory techniques. First, univariate time series data must be accompanied by a time-series data variable since two variables are required to execute

missing data theory techniques. Secondly, I noted that some techniques require the use of random number generators to utilize the multiple imputation-based techniques, such as multiple imputation using linear regression (MI-LR) in IBM SPSS 25. In addition, using the EM algorithm as a missing data theory technique or utilizing it to conduct Little's (2020) MCAR test to confirm the assumption that the data was missing completely at random required the use of a random number generator in IBM SPSS 25 (IBM knowledge center, 2021). As a result, I activated the Merzenne Twister random number generator with a random seed when it was required to go through the various portions of the intervention process.

For this intervention to occur at the missing value location, where the pretest value described as the "Original Numerical Value" was positioned in the data set, the following parameters were able to be manipulated to simulate the U.S. defense cost estimation missing data as planned in Chapter 3 for the pre-experiment (Aljuaid & Sasi, 2016; Strike et al., 2001):

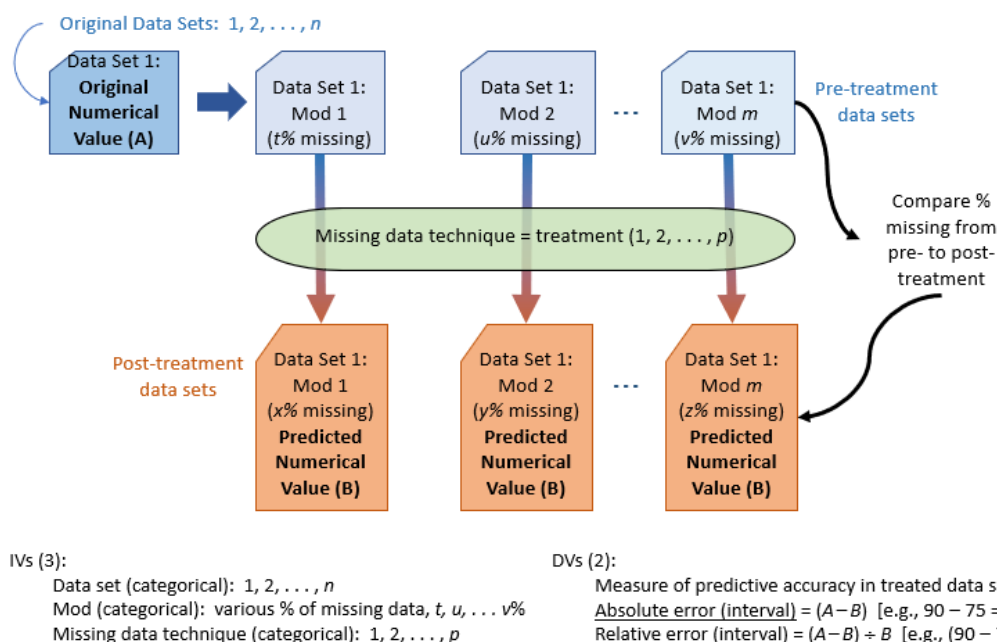
- The data set type selected from within the data matrix (independent variable 1).
- The percentage of missingness of the observations/programs with missing data (independent variable 2), based on the assumption that all missing software program observations were MCAR.
- The missing data theory technique (independent variable 3).

Eight different percentages of missing values were simulated per each of the missing data theory techniques applied (e.g., 5%, 10%, 15%, 20%, 25%, 30%, 35%,

40%). Since U.S. defense cost estimation data levels of analogous systems are generally not large data sets, this study did not exceed a missing data percentage rate beyond 40%. Figure 3 below is a full schematic of this study in which it defines the independent and dependent variable definitions, and as well as depicts how pretest and posttest approximation errors were calculated.

Figure 3

Statistical Method to Perform this One Group Pretest-Posttest Design



Each data set required three missing data theory treatments to be applied to fill in incomplete data 56 times, resulting in 4,704 ($3 \times 56 \times 28$) treatments in order to execute this study. Fifty-six treatments were derived because of randomly selecting which of the 30 programs would receive a treatment, as well as would be constrained to the percentage of missing programs needed at the eight levels of missingness (5%, 10%, 15%, 20%, 25%, 30%,

35%, and 40%) per data set. Figure 4 shows how software programs' data values were removed-at random based on their order in the data matrix.

Figure 4

Removed-at-Random-Data-Value Positions to Create the Artificially Induced Missing Data Problem at Eight Percentage Levels of Missingness

			% of 30 Raw Calculation	1.5	3	4.5	6	7.5	9	10.5	12
			# of Missing Values (R)	2	3	5	6	8	9	11	12
			% of Missing Data	5%	10%	15%	20%	25%	30%	35%	40%
Start With 50 Programs in Time Series Order	Selected 30 Analogous Software Programs	Updated Time Series/ Position Number	Assigned Pre-Experiment Program No.	Random Number Generator Set #1 Position	Random Number Generator Set #2 Position	Random Number Generator Set #3 Position	Random Number Generator Set #4 Position	Random Number Generator Set #5 Position	Random Number Generator Set #6 Position	Random Number Generator Set #7 Position	Random Number Generator Set #8 Position
1	3	1	P1						REMOVE	REMOVE	REMOVE
2	5	2	P2		REMOVE		REMOVE		REMOVE	REMOVE	REMOVE
3	6	3	P3	REMOVE		REMOVE	REMOVE			REMOVE	REMOVE
4	9	4	P4								
5	10	5	P5								
6	11	6	P6			REMOVE		REMOVE			
9	14	7	P7			REMOVE				REMOVE	
10	17	8	P8					REMOVE		REMOVE	
11	18	9	P9	REMOVE	REMOVE		REMOVE				
12	21 (COWBOY)	10	P10								
13	22	11	P11					REMOVE			
14	23	12	P12				REMOVE	REMOVE			
16	25	13	P13								
17	26	14	P14					REMOVE			
18	27	15	P15							REMOVE	REMOVE
20	28	16	P16						REMOVE		REMOVE
21 (COWBOY)	29	17	P17					REMOVE			REMOVE
22	30	18	P18		REMOVE		REMOVE		REMOVE	REMOVE	REMOVE
23	31	19	P19						REMOVE	REMOVE	REMOVE
24	32	20	P20								REMOVE
25	33	21	P21						REMOVE		
26	34	22	P22								
27	35	23	P23								
28	36	24	P24			REMOVE			REMOVE		
29	37	25	P25					REMOVE	REMOVE	REMOVE	REMOVE
30	42	26	P26								
31	43	27	P27					REMOVE			
32	44	28	P28							REMOVE	
33	46	29	P29				REMOVE			REMOVE	REMOVE
34	47	30	P30			REMOVE			REMOVE		REMOVE
...	thru 50										

Figure 5 shows how the methodology was applied to provide missing data theory treatments. As a result, each individual missing data theory treatments occur a total of 1,568 (56×28) each in this research study design.

Figure 5

Systematic Approach to Artificially Induce the Missing Data Problem at Eight Levels of Missingness for Three Missing Data Theory (MDT) Technique Treatments

Runs for Pre-Experiment per Data Set		Pre-Experimental Treatments/ Interventions to Conduct (Apply MDT Techniques to Randomly Removed Programs in Data Set)	Cases to Randomly Remove for Pre-Experiments	IV2	IV3	DV1	DV2
<u>Data Set Type (Categorical)</u>	<u>Runs for Pre-Experiment per Data Set Type's Values (# of Trials Per Each Data Set)</u>	<u>Random Number Generator Removes 56 Values to Create Missing Data Problem. Researcher Must Apply MDT Treatment for to following Missing Values (Total 56) Per Each Data Set</u>	<u>Number of Programs (Sample Size: n= 30)</u>	<u>Percentage of Missingness (Categorical)</u>	<u>MDT Techniques (Categorical)</u>	<u>Absolute Error (Numerical)</u>	<u>Relative Error (Numerical)</u>
	<u>Interventions to Treat Missing Values with MDT Techniques</u>	<u>56 Interventions = Label for Each Missing Values to Treat with MDT Techniques</u>	<u># of Programs</u>	<u>8 Levels</u>	<u>3 Levels -Traditional MDT Treatments</u>		
1	1	0.05_Program3_DataSet1	1	0.05	A. Listwise Delete		
2	2	0.05_Program9_DataSet1	2	0.10	B. Single Imputation		
3	3	0.10_Program2_DataSet1	3	0.15	C. Multiple Imputation		
4	4	0.10_Program9_DataSet1	4	0.20			
5	5	0.10_Program18_DataSet1	5	0.25			
6	6	0.15_Program3_DataSet1	6	0.30			
7	7	0.15_Program6_DataSet1	7	0.35			
8	8	0.15_Program7_DataSet1	8	0.40			
9	9	0.15_Program24_DataSet1	9				
10	10	0.15_Program30_DataSet1	10				
11	11	0.20_Program2_DataSet1	11				
12	12	0.20_Program3_DataSet1	12				
13	13	0.20_Program9_DataSet1	13				
14	14	0.20_Program12_DataSet1	14				
15	15	0.20_Program18_DataSet1	15				
16	16	0.20_Program29_DataSet1	16				
17	17	0.25_Program6_DataSet1	17				
18	18	0.25_Program8_DataSet1	18				
19	19	0.25_Program11_DataSet1	19				
20	20	0.25_Program12_DataSet1	20				
21	21	0.25_Program14_DataSet1	21				
22	22	0.25_Program17_DataSet1	22				
23	23	0.25_Program25_DataSet1	23				
24	24	0.25_Program27_DataSet1	24				
25	25	0.30_Program1_DataSet1	25				
26	26	0.30_Program2_DataSet1	26				
27	27	0.30_Program16_DataSet1	27				
28	28	0.30_Program18_DataSet1	28				
	29	0.30_Program19_DataSet1	29				
	30	0.30_Program21_DataSet1	30				
	31	0.30_Program24_DataSet1					
	32	0.30_Program25_DataSet1					
	33	0.30_Program30_DataSet1					
	34	0.35_Program1_DataSet1					
	35	0.35_Program2_DataSet1					
	36	0.35_Program3_DataSet1					
	37	0.35_Program7_DataSet1					
	38	0.35_Program8_DataSet1					
	39	0.35_Program15_DataSet1					
	40	0.35_Program18_DataSet1					
	41	0.35_Program19_DataSet1					
	42	0.35_Program25_DataSet1					
	43	0.35_Program28_DataSet1					
	44	0.35_Program29_DataSet1					
	45	0.40_Program1_DataSet1					
	46	0.40_Program2_DataSet1					
	47	0.40_Program3_DataSet1					
	48	0.40_Program15_DataSet1					
	49	0.40_Program16_DataSet1					
	50	0.40_Program17_DataSet1					
	51	0.40_Program18_DataSet1					
	52	0.40_Program19_DataSet1					
	53	0.40_Program20_DataSet1					
	54	0.40_Program25_DataSet1					
	55	0.40_Program29_DataSet1					
	56	0.40_Program30_DataSet1					

Study Results

Overall, my results support rejecting the null hypothesis for this research.

Traditional missing data theory techniques could mitigate the current gap in literature because it tested the level of predictive accuracy of three types of missing data theory techniques to improve the reliability and completeness of defense historical data. The empirical results from this data-driven research supports that missing data theory techniques can be taught in the Business—Cost Estimating discipline. In addition, it can be used whenever missing and incomplete values are present in a physical data matrix that a cost estimator is using to develop a cost estimate. The evaluative measures used to evaluate the study in more detail will be covered in the next two sections.

First Evaluation Measure to Determine Predictive Accuracy

The sample size for this pre-experimental research was comprised of 30 out of 50 analogous and synthetic software development programs and was able to inform the population of 30 DoD MAIS programs. As a generalized solution based on the testing and measurement of predictive accuracy, applying missing data theory treatments to DoD MAIS programs could be considered when needed. All 30 analogous synthetic software development programs had 28 software cost estimation types of data in which I was able to measure the predictive accuracy randomly across the same projects using eight percentage levels of missingness per data set type. Appendix A is where the results of each pre-experimental trial(run) per data set (Tables A1-A28) list each individual trial result for which missing data theory technique's calculated' "Predicted Numerical Value" came in closest to the "Original Numerical Value". Table 6 summarizes the 28 results by

showing which missing data theory technique had the closest predictive accuracy per each data set type.

Table 6

Summary of Closest Missing Data Theory (MDT) Technique Predictive Accuracy Results for Twenty-Eight Empirically Tested U.S. Defense Software Cost Estimating Data Types

Id	Experiment	Data Set Type	Closest MDT	Closest MDT
	Per Data Set		Technique for Predictive Accuracy	Technique Score out of 56 Experiments Per Data Set/ (By Percentage-Lowest Absolute and Relative Error Occurred)
1	DataSet1	Number of External Interface Requirements	MI-LR	46 (82%)
2	DataSet2	Initial SLOC – New	MI-LR	37 (66%)
3	DataSet3	Initial SLOC – Modified	SI-Mean	31 (55%)
4	DataSet4	Initial SLOC – Reused	SI -Mean	30 (54%)
5	DataSet5	Final SLOC – New	MI-LR	31 (55%)
6	DataSet6	Final SLOC – Modified	SI-Mean	29 (52%)
7	DataSet7	Final SLOC – Reused	SI -Mean	34 (61%)
8	DataSet8	DM % - Modified*	MI-LR	29 (52%)
9	DataSet9	CM % - Modified*	SI-Mean	31 (55%)
10	DataSet10	IM % - Modified*	SI -Mean	31 (55%)
11	DataSet11	DM % - Reused*	Both Perfect	56 (100%)

12	DataSet12	CM % - Reused*	Both Perfect	56 (100%)
13	DataSet13	IM % - Reused*	MI-LR	33 (59%)
14	DataSet14	Final Software Requirements Analysis Effort Hours	SI -Mean	30 (54%)
15	DataSet15	Final Software Architectural Design Effort Hours	MI-LR	33 (59%)
16	DataSet16	Final Software Detailed Design Effort Hours	SI -Mean	32 (57%)
17	DataSet17	Final Software Construction Effort Hours	SI-Mean	32 (57%)
18	DataSet18	Final Software Integration Effort Hours	SI -Mean	34 (61%)
19	DataSet19	Final Software Qualification Testing Effort Hours	SI -Mean	29 (52%)
20	DataSet20	Final Software Documentation Management Effort Hours	SI -Mean	32 (57%)
21	DataSet21	Final Software Configuration Management Effort Hours	SI -Mean	32 (57%)
22	DataSet22	Final Software Quality Assurance Effort Hours	SI -Mean	31 (55%)

23	DataSet23	Final Software Verification Effort Hours	MI-LR	33 (41%)
24	DataSet24	Final Software Validation Effort Hours	SI -Mean	35 (63%)
25	DataSet25	Final Software Review Effort Hours	SI -Mean	34 (61%)
26	DataSet26	Final Software Audit Effort Hours	Equal Predictive Accuracy	N/A
27	DataSet27	Final Software Problem Resolution Effort Hours	MI-LR	31 (45%)
28	DataSet28	Final Cybersecurity Effort Hours	SI -Mean	54 (96%)

The results show that out of the three missing data theories applied SI-Mean had the strongest level of predictive accuracy when experimental results were assessed at the individual data set level. Out of the 28 data sets results in Appendix A (Tables A1-A28), SI-Mean had a lower absolute and relative error in 16 data sets compared to only eight having the least amount of approximation error in MI-LR techniques.

When aggregating all summary results together, MI-LR produced “Predicted Values” that were within 20% of the “Original Numerical Value” 18.6% (292 out of 1,568) of the time when it was tested. Ironically, single imputation using the mean (SI-Mean) produced “Predicted Numerical Values” that were within 20% of the “Original Numerical Value” 16.4% (257 out of 1,568) of the time at the aggregate level. With this finding, at the aggregate level, MI-LR’s measure of predictive accuracy gets closer to the

“ground truth” or “Original Numerical Value” out of the three missing data theory techniques applied in this empirical study. Table 7 shows what degree the predictive accuracy of each missing data theory technique came close to the “Original Numerical Value”. Table 7 also answers this pre-experimental design’s answer to the RQ: To what degree can traditional missing data theory techniques accurately solve cost estimators’ and engineering managers' unreliable and incomplete data problem when data values are missing from a representative U.S. defense cost estimation data matrix?

Table 7

Degree to Which Missing Data Theory Techniques Can Solve the U.S. Cost Estimators’ Unreliable and Incomplete Data Problem Based on Approximation Error

Predictive Accuracy Results (Error)	MDT 1 LD	MDT 2 SI-Mean	MDT 3 MI-LR
Above 100% of Original Value	N/A	47.5%	48.7%
Within 80% or Less of Original Value	N/A	10.1%	8.9%
Within 60% or Less of Original Value	N/A	11.8%	10.9%
Within 40% or Less of Original Value	N/A	14.2%	12.9%
Within 20% or Less of Original Value	N/A	16.4%	18.6%

Second Evaluation Measure to Test Main Effects & Interactions

In addition to measuring the level of predictive accuracy, I measured the main effects and interactions between the independent and dependent variables by conducting ANOVA testing. A two-way repeated measures ANOVA was conducted to explain the interaction of this study’s two dependent variables using the following null and alternate hypotheses:

H_01 : There are no significant differences evident between the data sets’ mean absolute and mean relative error of actual values “Original Numerical Values” in

comparison to those that are computed “Predicted Numerical Values” using missing data theory 1, the Complete Case Analysis/ Listwise Delete approach?

H_{a1} : There are significant differences evident between the data sets’ mean absolute and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 1, Complete Case Analysis/ Listwise Delete approach? The means are not equal.

H_{02} : There are no significant differences evident between the data set’s mean absolute and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 2, a Single Imputation approach?

H_{a2} : There are significant differences evident between the data sets’ mean absolute and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 2, a Single Imputation approach? The means are not equal.

H_{03} : There are no significant differences evident in the data sets’ mean absolute error and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values” using missing data theory 3, the Multiple Imputation approach?

H_{a3} : There are significant differences evident between the data sets’ mean absolute error and mean relative error of actual values “Original Numerical Values” in comparison to those that are computed “Predicted Numerical Values”

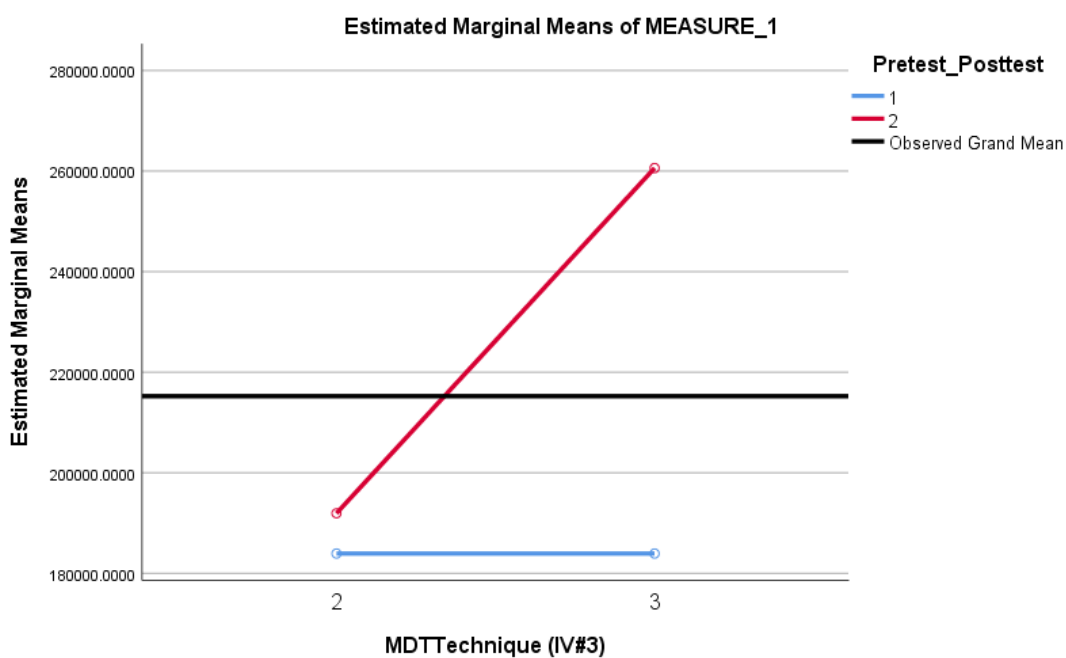
using missing data theory 3, the Multiple Imputation approach? The means are not equal.

Each null and alternate hypothesis was the same for all three hypotheses to determine if the means are equal between the actual values “Original Numerical Values” and the computed “Predicted Numerical Values”. For missing data theory 1, Complete Case Analysis/ Listwise Delete (LD), the results from the treatment required the incomplete variable to be dropped. As a result, a statistical analysis was unable to be ran with data not present. Unfortunately, the first hypothesis could not be assessed.

The results for missing data theory 2, Simple Imputation (SI-Mean) and missing data theory 3, Multiple Imputation (MI-LR) were analyzed in IBM SPSS 25 to better understand the variable behavior and to test for significance. The estimated marginal means chart provides a graphical illustration that the means are not equal. Figure 6 would depict if an interaction between the different means across the experiment under both SI-Mean as the horizontal blue, and MI-LR as the horizontal red line. No interaction occurred between the two dependent variables.

Figure 6

Plot of the Results to Assess Interaction Between Approximation Error (Dependent Variables) of the Actual/Pretest (1) and Computed Posttest (2) Value Means



Continuing with the second evaluative measure, I then assessed the three independent variables for an interaction as well. Figure 7 depicted that for the F statistic, only the data set type (IV1), was significant using a p value of $< .05$. This means that there were no statistically significant outputs to determine if there was an interaction between the three independent variables. Figure 7 shows the between-subject effects results which provide the visibility to determine if there was an interaction between the independent variables used within this study's pre-experimental design.

Figure 7*Interaction Analysis for the Two-Way Repeated Measures ANOVA*

Tests of Between-Subjects Effects						
Measure: MEASURE_1						
Transformed Variable: Average						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	1.872E+14	1	1.872E+14	383.284	.000	.125
ProgramTreatmentofMissingnessIV#1	2.961E+12	7	4.230E+11	.866	.533	.002
DataSetAllTypeIV#2	9.673E+14	27	3.583E+13	73.338	.000	.424
MDTTechniqueIV#3	1.312E+12	1	1.312E+12	2.685	.101	.001
ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2	2.116E+13	189	1.119E+11	.229	1.000	.016
ProgramTreatmentofMissingnessIV#1 * MDTTechniqueIV#3	4.262E+11	7	6.089E+10	.125	.997	.000
DataSetAllTypeIV#2 * MDTTechniqueIV#3	8.720E+12	27	3.230E+11	.661	.907	.007
ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2 * MDTTechniqueIV#3	6.662E+12	189	3.525E+10	.072	1.000	.005
Error	1.313E+15	2688	4.885E+11			

Lastly, to confirm if there were main effects, I assessed the within subjects effects, and determined that there were significant effects on three sources due to using a $p < .05$ for the following:

- Pretest_Posttest (for the “Original Value” and “Predicted Value”)
- Pretest_Posttest * Program Treatment/Missingness (IV1)
- Pretest_Posttest * Program Treatment/Missingness (IV1) * DataSetAllTypes (IV2)

Figure 8 displays where the F statistics demonstrated to me that a main effect does exist with an effect of one independent variable on the dependent variable, in this case the

outcome variables for this pre-experiment. More importantly, this test also revealed to me that there were statistically significant differences in mean absolute error and relative error of the pretest and posttest value in this pre-experimental design since the p value < .05. With .016 being less than .05 we reject the null. If there is less than a 5% chance of a result as extreme as this sample and the null hypothesis were true, the null hypothesis is rejected. Figure 8 depicts that the means are not equal, and that we should reject the null hypothesis for this research.

Figure 8

Statistical Significance and Main Effect from the Two-Way Repeated Measures ANOVA

Tests of Within-Subjects Effects							
Measure: MEASURE_1							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Pretest_Posttest	Sphericity Assumed	1.994E+12	1	1.994E+12	5.763	.016	.002
	Greenhouse-Geisser	1.994E+12	1.000	1.994E+12	5.763	.016	.002
	Huynh-Feldt	1.994E+12	1.000	1.994E+12	5.763	.016	.002
	Lower-bound	1.994E+12	1.000	1.994E+12	5.763	.016	.002
Pretest_Posttest * ProgramTreatmentofMiss ingnessIV#1	Sphericity Assumed	1.102E+13	7	1.575E+12	4.550	.000	.012
	Greenhouse-Geisser	1.102E+13	7.000	1.575E+12	4.550	.000	.012
	Huynh-Feldt	1.102E+13	7.000	1.575E+12	4.550	.000	.012
	Lower-bound	1.102E+13	7.000	1.575E+12	4.550	.000	.012
Pretest_Posttest * DataSetAllTypeIV#2	Sphericity Assumed	1.236E+13	27	4.579E+11	1.323	.123	.013
	Greenhouse-Geisser	1.236E+13	27.000	4.579E+11	1.323	.123	.013
	Huynh-Feldt	1.236E+13	27.000	4.579E+11	1.323	.123	.013
	Lower-bound	1.236E+13	27.000	4.579E+11	1.323	.123	.013
Pretest_Posttest * MDTTechniqueIV#3	Sphericity Assumed	1.312E+12	1	1.312E+12	3.791	.052	.001
	Greenhouse-Geisser	1.312E+12	1.000	1.312E+12	3.791	.052	.001
	Huynh-Feldt	1.312E+12	1.000	1.312E+12	3.791	.052	.001
	Lower-bound	1.312E+12	1.000	1.312E+12	3.791	.052	.001
Pretest_Posttest * ProgramTreatmentofMiss ingnessIV#1 * DataSetAllTypeIV#2	Sphericity Assumed	1.212E+14	189	6.412E+11	1.853	.000	.115
	Greenhouse-Geisser	1.212E+14	189.000	6.412E+11	1.853	.000	.115
	Huynh-Feldt	1.212E+14	189.000	6.412E+11	1.853	.000	.115
	Lower-bound	1.212E+14	189.000	6.412E+11	1.853	.000	.115
Pretest_Posttest * ProgramTreatmentofMiss ingnessIV#1 * MDTTechniqueIV#3	Sphericity Assumed	4.262E+11	7	6.089E+10	.176	.990	.000
	Greenhouse-Geisser	4.262E+11	7.000	6.089E+10	.176	.990	.000
	Huynh-Feldt	4.262E+11	7.000	6.089E+10	.176	.990	.000
	Lower-bound	4.262E+11	7.000	6.089E+10	.176	.990	.000
Pretest_Posttest * DataSetAllTypeIV#2 * MDTTechniqueIV#3	Sphericity Assumed	8.720E+12	27	3.230E+11	.933	.563	.009
	Greenhouse-Geisser	8.720E+12	27.000	3.230E+11	.933	.563	.009
	Huynh-Feldt	8.720E+12	27.000	3.230E+11	.933	.563	.009
	Lower-bound	8.720E+12	27.000	3.230E+11	.933	.563	.009
Pretest_Posttest * ProgramTreatmentofMiss ingnessIV#1 * DataSetAllTypeIV#2 * MDTTechniqueIV#3	Sphericity Assumed	6.662E+12	189	3.525E+10	.102	1.000	.007
	Greenhouse-Geisser	6.662E+12	189.000	3.525E+10	.102	1.000	.007
	Huynh-Feldt	6.662E+12	189.000	3.525E+10	.102	1.000	.007
	Lower-bound	6.662E+12	189.000	3.525E+10	.102	1.000	.007
Error(Pretest_Posttest)	Sphericity Assumed	9.301E+14	2688	3.460E+11			
	Greenhouse-Geisser	9.301E+14	2688.000	3.460E+11			
	Huynh-Feldt	9.301E+14	2688.000	3.460E+11			
	Lower-bound	9.301E+14	2688.000	3.460E+11			

Post hoc tests were not performed for missing data theory techniques (independent variable 3) because there are fewer than three groups that made it into my IBM SPSS 25

analysis. Mauchly's test for sphericity was not required because there were less than three dependent variables that could be ran for the repeated measures ANOVA.

To reiterate, the purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. In summary, the pre-experimental findings and results empirically demonstrate that missing data theory techniques could be a viable option to correct imperfect data that is unreliable or incomplete with a data value that is closer to the ground truth of the original numerical values.

Summary

Based on the results from the 4,704 treatments, I was able to empirically measure the predictive accuracy of three missing data theory techniques: complete case analysis using listwise delete (LD), single imputation using the mean (SI-Mean), and multiple imputation using linear regression (MI-LR). To answer the study's RQ, the results from using various percentages levels of missingness and 28 different data set types supported that at least two out of the three missing data theory techniques can render a representative U.S. cost estimation data matrix more complete when data values are missing. Specifically, missing data theory technique 2, single imputation's "Predicted Numerical Value" rendered a forecasted value that was closest to the "Original Numerical Value" when tallying results of all experimental runs at the data set level of

only 56 cases each to assess. Single imputation using the mean (SI-Mean) had the lowest absolute and relative error 52.3% of the time and multiple imputation using linear regression (MI-LR) had the next lowest absolute and relative error of 47.7% when analyzing which missing data theory technique had the closest predictive accuracy within each data set; limited to only 56 treatments.

At the aggregate level, when comparing 1,568 (56*28) empirical results for all 28 data sets combined per each missing data theory technique treatment, multiple imputation using linear regression (MI-LR) had the closest predictive accuracy when compared to single imputation using the mean (SI-Mean) and listwise delete (LD). Multiple imputation (MI-LR) produced “Predicted Values” that were within 20% of the “Original Value” 18.6% of the time when it was used as a treatment. Ironically, single imputation using the mean (SI-Mean) only produced “Predicted Values” that were within 20% of the “Original Value” 16.4% of this time at the aggregate level. With this finding, it appears that at the aggregate level, Multiple Imputation measure of predictive accuracy to actual values gets closer to the “ground truth” true value.

Unfortunately, the complete case analysis using listwise delete (LD) did not produce any forecasted “Predicted Numerical Value” to assess its predictive accuracy since the execution of this traditional approach is to drop values that do not have any data. As a result, this technique did not render a “Predictive Value” to measure listwise delete’s predictive accuracy on a representative U.S. defense cost estimation data matrix. As an observation of the treatment, it does not forecast any value and thus is not a

feasible method to make unreliable and incomplete data in Business—Cost Estimating complete.

Chapter 5 will provide the conclusions from this final study. It will also include recommendations and further studies that could be continued as a result of this research. The conclusions, limitations, and recommendations are clearly described for the scope of this study, and describe how the integration of this study fits into the state of knowledge described in the researched literature review in order to close a gap for the Business—Cost Estimating discipline.

Chapter 5: Discussion, Conclusions, and Recommendations

The purpose of this quantitative study was to test and measure the level of predictive accuracy of missing data theory techniques that are referenced as traditional approaches in the literature, compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline and determine which theories render incomplete and missing data sets in a single data matrix most reliable and complete under several missing value percentages. The nature of this study was a quantitative method approach to inquiry using a pre-experimental study design. Various experimental study designs (pre-, quasi-, or true experiments) are a proven approach to comparatively test and measure the predictive accuracy of missing data theory techniques using a pretest-posttest no control group design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Crammer, 2018; Kirk, 2013; Reichardt, 2019; Shadish et al., 2002; Shek & Zhu, 2018; Singleton & Strait, 2010). To elucidate how effective each missing data theory technique was, a publicly sourced nonproprietary data matrix was obtained and manipulated to experiment on 28 out of 34 ratio scale/numerical software cost estimation data set types (independent variable 2) used within the U.S. defense cost estimating discipline from the representative data matrix. In addition, eight levels of missing data percentages (independent variable 1) were assessed across each data set type to compare the measures of predictive accuracy, for each of the three missing data theory techniques (independent variable 3). Once the data sets were exported to a flat file in Microsoft Excel, the experiment followed a four-step process, like the research conducted by Idri et al. (2016b). The actual known data values (pretest values), provided the pretest baseline,

was used to compare how accurately each missing data theory techniques produced its respective “Predicted Numerical Value” (posttest value). The “Original Numerical Values” (pretest /priori values) were removed-at-random to create missing values within the data matrix by using a random number generator (Idri et al., 2015a, 2016a, 2016b; Idri et al., 2015b, 2016c; Research Randomizer, 2020). Next, the complete data set generation occurred in which the missing data theory technique (independent variable 3) treatment variables were then calculated and applied to make each of the 28 data sets complete again. After which, the measurement of predictive accuracy evaluation began, and measured the outcome variables, the error approximation values, by calculating the absolute error and relative error values between the pretest and posttest values from the pre-experiment. A two-way repeated measures ANOVA was used to test the study’s null and alternative hypotheses, and to determine if there was a significant interaction between independent variables.

Interpretation of Findings

The key finding was that out of the three missing data theories applied SI-Mean had the strongest level of predictive accuracy when experimental results were assessed at the individual data set level. Out of the 28 data sets results (Tables A1-A28), SI-Mean had a lower absolute and relative error in 16 data sets compared to only eight having the least amount of approximation error in MI-LR techniques. Considering many studies before me have acknowledged that multiple imputation has better prediction accuracy. Both techniques performed equally well on data sets 11, 12, and 26.

When looking at each technique in isolation of each other, the key finding showed that multiple imputation results had a closer prediction accuracy than simple imputation when it was closer to the “Original Numerical Value”. Multiple imputation was also able to calculate a value that was within 20% of the “Original Numerical Value(s)” across all data sets 18.6% of the time (292 out of 1,568 multiple imputation treatments). In comparison, single imputation was able to predict within 20% of the “Original Values” 16.7% of the time (257 out of 1,568 single imputation treatments). This tells us that whenever multiple imputation had the closest predictive accuracy, even though not as many times at the data set level, it tended to be within 20% of the “Original Value” when it was close.

Overall, these results support rejecting the null hypothesis for this research. Traditional missing data theory techniques could mitigate the current gap in literature because it tested the level of predictive accuracy of three types of missing data theory techniques to improve the reliability and completeness of defense historical data. The empirical results from this data-driven research support that missing data theory techniques could be taught in the Business—Cost Estimating discipline. In addition, it could be used whenever missing and incomplete values are present in a physical data matrix that a cost estimator is using to develop a cost estimate.

Limitations of the Study

The research design of this study was limited based on the instrumentation selected to test predictive accuracy. I used IBM SPSS 25 as the instrumentation to conduct a pre-experimental design to test the predictive accuracy of missing data theory

techniques on a representative U.S. defense cost estimating data matrix. SPSS is recognized in the academic community and has the statistical capability and processing power to assess data that has incomplete and missing values (Enders, 2010). I leveraged the statistical analysis capability that is provided in the Missing Value Analysis module, Multiple Imputation functionality of IBM SPSS 25. The Missing Value Analysis module and Multiple Imputation functionality in IBM SPSS 25 has the computational ability to compute traditional missing data theory algorithms. As a result of this functionality, IBM SPSS 25 was applied as the instrumentation for this inaugural study that tested missing data theoretical techniques' predictive accuracy when applied to the U.S. defense cost estimation domain. Despite this being a limitation of this study, treatments were replicated and assessed as a one group pretest-posttest no control group/pre-experimental design intervention.

Not having a control group for the one group pretest-posttest pre-experimental research design was a weakness. However, it was not pertinent to have a control group to answer this RQ because it focused on "Original Numerical Values" as a test group, vice testing groups that could have been comprised of human subjects that are exposed to outside experimental factors that may skew results. In social work, for example, human subjects under intervention studies make it difficult to control for outside influences that often skew responses that may not be isolated, and thus require a control group to compare results (Thyer, 2012). The use of data as the subject in this intervention using a one group pretest-posttest design enabled me to minimize potential threats to internal and external validity. Each independent variable completely controlled how I manipulated the

pre-experiments for each data set to only receive three types of treatments and were evaluated within the confines of this intervention study's independent variables. I was able to mitigate any confounding or extraneous variables from entering the intervention study, each dependent variable was instantly evaluated within a short time-box to answer this study's RQ after the intervention.

In addition, the construct of this study remained strong because of its well-defined and focused scope to test and measure the level of predictive accuracy of missing data theory as it pertains to (a) listwise deletion (LD) or complete case analysis, (b) single imputation, and (c) multiple imputation on an IRB approved and representative U.S. defense cost estimation data matrix. This narrowed focus was not biased but was intentional in order to address the specific research questions of this study that took a first look at applying traditional missing data theory to the U.S. defense cost estimation domain, something that, according to found literature, has not previously been done. Further studies can extend the scope of this study and add to the literature to expand outcomes of this analysis.

Recommendations

Artificially inducing the missing values into the data matrix was chosen because it was an approach that has been adopted in several missing data experimental studies, and allowed me the aptitude to apply several missing data theory technique treatments to a sample, and test for predictive accuracy of this intervention while adding knowledge and understanding to the gap in the literature (Brown & White, 2017; DAU, 2018a; GAO, 1972, 2009; Hill, 2011; Idri et al., 2015a, 2016a, 2016b; Idri et al., 2015b, 2016c; Mislick

& Nussbaum, 2015; Twala et al., 2006). In being able to apply missing data theory techniques to the current challenges of not having reliable and complete data always available, filling in this gap by incorporating the techniques empirically presented is strongly recommended to the discipline as an additional option to improve data quality and mitigating the unreliable and incomplete data problem in the Business—Cost Estimating discipline.

Since this study only empirically tested traditional missing data theory techniques, additional statistical learning, machine learning, and other imputation and model-based techniques should be tested to further explore this gap that has been untapped since 1972 (DAU, 2018a; GAO, 1972, 2009, 2020; Mislick & Nussbaum, 2015). In addition, this study design leveraged a complete data set to allow for the “Predicted Numerical Value(s)” from each data matrix to be assessed against each “Original Numerical Value(s)” as provided from a nonproprietary data matrix. The data matrix contained data sets that were representative of what could be found in databases used by cost estimators, engineering economists, and engineering managers within the defense cost estimating discipline (e.g., from FACADE, USASpending.gov (2021), IT Dashboarddata.gov (2021), etc.). In future studies, additional other representative cost data matrices and data set types could be explored that extend beyond the United States., as well as beyond software cost estimation.

Implications

This study was important to conduct because “reliable and comprehensive cost data is essential to produce credible cost estimates as required in both (policy) statute and

regulation” (Morin, 2017, p. 1). Brown and White (2017) agreed with Morin and reported that the federal defense department lacked the data, both in volume and quality, needed to conduct effective cost estimates. Together, these authors acknowledged that cost estimate realism was and still is essential and needed to support engineering and program managers to gain the authority and approvals needed to proceed into the development and contractual procurement of critical engineering systems. This study offered a different perspective on an established problem on what hands-on-treatment-options can be used when historical databases or other data resources contain substantial amounts of missing data (Strike et al., 2001). Conducting research to “improve analyst productivity, quality of cost estimates, close data gaps, and provide the cost acquisition, and resource allocation organizations with data required for better analysis and decision-making” is significant (Morin, 2017, p. 1).

Significance to Theory

The outcome of this study may offer defense industry cost estimators, engineering economists, engineering managers, defense cost estimating repository database administrators, and possibly data scientists with an objective option in how to deal with missing, incomplete, or unreliable data values when they appear within a data matrix. Applying and continuing to test missing data theory on actual complete data sets that are relevant to the problem could provide the empirical evidence needed to prove or disprove how well various missing data theories are able to fill missing data value gaps. Contingent on the outcomes observed after randomly removing variables to simulate a missing data problem, this could improve the missing, incomplete, and unreliable data

problem that is experienced within the U.S. defense cost estimation discipline. In addition, U.S. defense cost estimators tend to build models with small data matrices, n less than or equal to 30, in which this empirical study that tested the performance of small sample size data sets, and how well missing data theories' predictive accuracy levels were explored could be incorporated into future government documents and textbooks.

Significance to Practice

Cost estimators of defense weapon systems must have access to reliable and complete data sets from the historical database repositories and other sources they access to develop accurate engineering economic requirements. Cost estimates, the end-product from cost estimating, is a critical document needed to request the right amount of budget authority from Congress to fund any future investments (Mislick & Nussbaum, 2015). When databases have null values, obvious errors, and blank cells because of various systemic data problems, it is up to the cost estimator to make the decision as to how to use this type of data value within a data matrix to feed a cost estimate element. In layman's terms, there is no disciplined approach taught to defense cost estimators in what data values to use or not use in their physical data matrix when missing, incomplete, or unreliable data values appears (DAU, 2018a). With over 250 defense cost estimators within the Business—Cost Estimating career field, there is no established standard as to how to handle this problem within the defense cost estimating discipline (DAU, 2018a; DAU, 2018b). Because of this study's empirical results, teaching engineering managers and cost estimators within the discipline about single and multiple imputation as feasible options to handle missing, incomplete, or unreliable data values, could reduce the number

of flawed cost estimates that lead to program cost overruns and unplanned additional federal budget request.

Significance to Social Change

Accurately forecasting estimates for engineering requirements could save projects and programs from growing cost overruns and improve U.S. federal planning decisions (Christensen, 1993; Christensen & Gordon, 1998; Deloitte, 2016; Saeed et al., 2018). In addition, positive social change could be realized by improving the current techniques cost estimators and engineering managers use to produce and provide more accurate, reliable, and credible cost estimates to federal decision-makers. Moreover, research that could advance cost data quality and improvement efforts could also increase the amount of historical DoD cost data that can be used in analyses. Overall, a new way of doing business to use single and multiple imputation may save cost estimator's, engineering economists', engineering manager's and database administrator's valuable time by using a newly proven technique to improve data in a shorter amount of time. In turn, this contribution to the cost estimation discipline has the potential to reduce the cost of an estimator's research time and reduce the cost required to collect additional data.

Conclusions

To further the application of missing data theory to the U.S. defense cost estimation discipline, I modeled the missing data problem by simulating the conditions that defense industry cost estimators, engineering economists, engineering managers, and defense cost estimating repository database administrators experience. I used the pre-experimental study design to apply a missing data theory complete case analysis

treatment and two statistical based missing data theory treatments on the same group of randomly selected data from a complete U.S. represented cost estimation data matrix. Complete case analysis did not render data sets more complete; however, single and multiple imputations did restore data sets with values that were within 20% of its “Original Numerical Value(s)” 16.4% and 18.6% of the experiments, respectively. The data matrix contained an appropriate required sample size of 30 DoD software programs and contained 28 data sets to test missing data theory techniques on. The “Predicted Numerical Value”, as determined by each missing data theory technique, served as the posttest value in this experiment and helped calculate the study’s dependent variables, described as its measures of predictive accuracy. Stated differently, the two dependent variables that captured the predictive accuracy for this study were absolute error and relative error. The absolute errors and relative errors were calculated from pretest and posttest values.

To answer the RQ, I used the pre-experimental research design of the one group pretest-posttest no control group design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Crammer, 2018; Reichardt, 2019; Shadish et al., 2002; Shek & Zhu, 2018; Singleton & Strait, 2010; Thyer, 2012). Significance testing was performed by conducting a two-way repeated measures ANOVA. The estimated marginal means plot, and the *F* statistic were used to test the main effects and interaction between variables.

This quantitative research method of inquiry helped determine how well defense cost estimators could handle historical data sets with the use of missing data theory techniques (Crammer, 2018; Kirk, 2013; Shek & Zhu, 2018; Thyer, 2012). By randomly

removing data values from a complete data set, an empirical examination of new data values was quantitatively created, assessed, and proved that both single imputation and multiple imputation missing data theory techniques have the ability to improve the unreliable and incomplete data quality problem that is currently experienced in the U.S. defense cost estimation discipline of Business—Cost Estimating.

By conducting this empirical research, I closed a gap in the U.S. cost estimation discipline and added to the research, knowledge, and understanding which serve as rationale for employing additional options for cost estimators to perform more reliable and complete cost estimation products. Major DoD engineering-based acquisition projects and programs cost estimates require reliable and complete data to forecast cost more accurately. The results of this empirical research could provide U.S. defense cost estimators with an evaluation of which one out of three missing data theory techniques could serve as a hands-on-treatment-options that could handle the unreliable and incomplete data problem when building a cost estimate (DAU, 2018a; GAO, 2009, 2020; Morin, 2017).

Based on the gap discovered in the literature review, this study was necessary because it could potentially improve data quality with missing data theory techniques that support the need for “reliable and comprehensive cost data ...to produce credible cost estimates as required in both (policy) statute and regulation” (Morin, 2017, p. 1). In the past, GAO (1972) reported that the federal defense department lacked the data, both in volume and quality, needed to conduct effective cost estimates (Brown & White, 2017). As a result, cost estimate realism to support future engineering systems’ (e.g., developed

software, aircraft, ships, business systems, autonomous systems, artificial intelligent systems, etc.) success is threatened. This study now provides a different perspective to address this established problem that historical databases contain substantial amounts of missing data (Strike et al., 2001) and could be potentially mitigated by a hands-on-treatment. By helping cost estimators, engineering managers, and database administrators in the federal defense department “improve analyst productivity, quality of cost estimates, close data gaps, and provide the cost, acquisition, and resource allocation organizations with data required for better analysis and decision-making”, an improvement to fund programs to an improved and more accurate estimated planned amount to complete an engineering project could be accomplished by now training cost estimators to use missing data theory techniques (Morin, 2017, p. 1).

References

- Abnane, I., & Idri, A. (2018). Improved analogy-based effort estimation with incomplete mixed data. In M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Annals of Computer Science and Information Systems*, 15, 1015-1024. <https://www.doi.org/10.15439/2018F95>
- Aittokallio, T. (2009). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2), 253-264. <https://www.doi.org/10.1093/bib/bbp059>
- Aljuaid, T., & Sasi, S. (2016, August). Proper imputation techniques for missing values in data sets. In *2016 International Conference on Data Science and Engineering* (pp. 1-5). IEEE. <https://www.doi.org/10.1109/ICDSE.2016.7823957>
- Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, 28(3), 301-309. <https://www.doi.org/10.1177/0049124100028003003>
- Allison, P. D. (2002). *Quantitative applications in the social sciences: missing data*. Sage. <https://www.doi.org/10.4135/9781412985079>
- Allison, P. D. (2010). *Missing data*. Sage.
- azzahra Amazal, F., Idri, A., & Abran, A. (2014). Software development effort estimation using classical and fuzzy analogy: a cross-validation comparative study. *International Journal of Computational Intelligence and*

Applications, 13(03), 1450013.1 – 1450013.19.

<https://www.doi.org/10.1142/S1469026814500138>

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses.

Journal of School Psychology, 48(1), 5-37.

<https://www.doi.org/10.1016/j.jsp.2009.10.001>

Blankers, M., Koeter, M. W., & Schippers, G. M. (2010). Missing data approaches in eHealth research: simulation study and a tutorial for nonmathematically inclined researchers. *Journal of Medical Internet Research*, 12(5).

<https://www.doi.org/10.2196/jmir.1448>

Boehm, B. W. (1981). *Software engineering economics* (Vol. 197). Prentice-hall.

Boehm, B. W. (1984). Software engineering economics. *IEEE Transactions on Software Engineering*, (1), 4-21.

Boehm, B. W. (2002). Software engineering economics. In M. Broy & E. Denert (Eds.)

Software pioneers (pp. 641-686). Springer. https://www.doi.org/10.1007/978-3-642-59412-0_38

Briand, L. C., Langley, T., & Wieczorek, I. (2000, June). A replicated assessment and comparison of common software cost modeling techniques. In *Proceedings of the 22nd international conference on Software engineering* (pp. 377-386). ACM.

<https://www.doi.org/0.1145/337180.337223>

Brown, G. E., & White, E. D. (2017). *An investigation of nonparametric DATA MINING*

TECHNIQUES for acquisition cost estimating. Defense Acquisition Research

Journal: A Publication of The Defense Acquisition University, 24(2), 302-332.

<https://www.doi.org/10.22594/dau.16-756.24.02>

Burke, R. P., & Spruill, N. L. (2016, April 15) Implementation memo to add a core certification course for the Business—Cost Estimating career field.

[https://www.dau.edu/training/career-](https://www.dau.edu/training/career-development/bce/Documents/BCF%20250%20Implementation%20Memo%20-%20signed%204-15-2016.pdf?listid=789a03cd-2f52-497c-a9c7-ebb4028426b5&preview=1&itemid=2)

[development/bce/Documents/BCF%20250%20Implementation%20Memo%20-](https://www.dau.edu/training/career-development/bce/Documents/BCF%20250%20Implementation%20Memo%20-%20signed%204-15-2016.pdf?listid=789a03cd-2f52-497c-a9c7-ebb4028426b5&preview=1&itemid=2)

[%20signed%204-15-2016.pdf?listid=789a03cd-2f52-497c-a9c7-](https://www.dau.edu/training/career-development/bce/Documents/BCF%20250%20Implementation%20Memo%20-%20signed%204-15-2016.pdf?listid=789a03cd-2f52-497c-a9c7-ebb4028426b5&preview=1&itemid=2)

[ebb4028426b5&preview=1&itemid=2](https://www.dau.edu/training/career-development/bce/Documents/BCF%20250%20Implementation%20Memo%20-%20signed%204-15-2016.pdf?listid=789a03cd-2f52-497c-a9c7-ebb4028426b5&preview=1&itemid=2)

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of Research on Teaching*, 171-246.

<http://jwilson.coe.uga.edu/EMAT7050/articles/CampbellStanley.pdf>

Carter, A. (2020). Better buying power: Guidance for obtaining greater efficiency and productivity in defense spending.

https://www.acq.osd.mil/fo/docs/USD_ATL_Guidance_Memo_September_14_2010_FINAL.PDF

Cartwright, M. H., Shepperd, M. J., & Song, Q. (2003, September). Dealing with missing software project data. In *Software Metrics Symposium, 2003. Proceedings. Ninth International* (pp. 154-165). IEEE.

<https://www.doi.org/10.1109/METRIC.2003.1232464>

Chen, X., Wei, Z., Li, Z., Liang, J., Cai, Y., & Zhang, B. (2017). Ensemble correlation-based low-rank matrix completion with applications to traffic data

imputation. *Knowledge-Based Systems*, 132, 249-262.

<https://www.doi.org/10.1016/j.knosys.2017.06.010>

Christensen, D. S. (1993). An analysis of cost overruns on defense acquisition contracts.

Project Management Journal, 24(3), 43–48.

<https://ntrs.nasa.gov/archive/nasa/cami.ntrs.nasa.gov/19950014332.pdf>

Christensen, D. S., & Gordon, J. A. (1998). Does a rubber baseline guarantee cost

overruns on defense acquisition contracts? *Project Management Journal*, 29(3),

43-51. <https://www.doi.org/10.1177/875697289802900307>

Conte, S. D., Dunsmore, H. E., & Shen, V. Y. (1986). *Software engineering metrics and models*. Benjamin-Cummings Publishing.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Houghton Mifflin.

Cranmer, G. A. (2018). One-group pretest-posttest design. In *The Sage encyclopedia of communication research methods* (Vols. 1-4) (pp. 1125-1126).

SAGE Publications. <https://www.doi.org/10.4135/9781483381411>

Dabkowski, M., & Valerdi, R. (2016). Blockmodeling and the estimation of evolutionary architectural growth in major defense acquisition programs. US Army.

<https://apps.dtic.mil/dtic/tr/fulltext/u2/1016757.pdf>

Defense Acquisition University. (2018a). *Business: Cost estimating courses*.

<https://www.dau.mil/cop/ce/Pages/Course.aspx>

Defense Acquisition University. (2018b). *Policy for business: Cost estimating*.

<https://www.dau.mil/policy#Business%20Cost%20Estimating|All|All|All|recent>

Defense Acquisition University. (2020). *Course Material Login Account*.

[https://myclass.dau.edu/webapps/blackboard/content/listContentEditable.jsp?content_id= 1282460_1&course_id= 92810881_1](https://myclass.dau.edu/webapps/blackboard/content/listContentEditable.jsp?content_id=1282460_1&course_id=92810881_1)

De la Garza, J. M., & Rouhana, K. G. (1995). Neural networks versus parameter-based applications in cost. *Cost Engineering*, 37(2), 14.

de Leeuw, E. D. (2001). Reducing missing data in surveys: An overview of methods. *Quality and Quantity*, 35(2), 147-160.

<https://www.doi.org/10.1023/A:1010395805406>

Deloitte. (2016). Cost overruns persist in major defense programs.

<https://www.prnewswire.com/news-releases/deloitte-study-cost-overruns-persist-in-major-defense-programs-300349671.html>

Department of Defense Instruction. (2017). Cost Analysis Guidance and

Procedures.<https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500073p.pdf>.

Department of Defense Manual. (2011). Cost and Software Data Reporting Manual.

<https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/500004m1.pdf>

Desai, D. J., Jain, T. S., Dwivedi, A. A., & Attar, A. D. (2016). Engineering economics and life cycle cost analysis.. *International Journal of Research in Engineering and Technology*, 5(03), 390-394 <http://tiny.mitre.org/A0F8>

Eirola, E., Doquire, G., Verleysen, M., & Lendasse, A. (2013). Distance estimation in numerical data sets with missing values. *Information Sciences*, 240, 115-128.

- Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.
<https://www.doi.org/10.1016/j.neucom.2013.07.050>
- Farr, J. V., & Faber, I. (2018). *Engineering Economics of Life Cycle Cost Analysis*. CRC Press. <https://www.doi.org/10.1201/9780429466304>
- Federal Procurement Data System – Next Generation (FPDS-NG). (2018).
https://www.fpds.gov/fpdsng_cms/index.php/en/
- Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage Publication.
- Fraser, N. M. & Jewkes, E. M. (2013). *Engineering economics: Financial decision making for engineers* (5th ed.) Pearson Canada.
- García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2), 263-282. <https://www.doi.org/10.1007/s00521-009-0295-6>
- García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2013). Classifying patterns with missing values using multi-task learning perceptrons. *Expert Systems with Applications*, 40(4), 1333-1341.
<https://www.doi.org/10.1016/j.eswa.2012.08.057>
- Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89, 52-65.
<https://www.doi.org/10.1016/j.eswa.2017.07.026>

- Garvey, P. R., Book, S. A., & Covert, R. P. (2016). *Probability methods for cost uncertainty analysis: A systems engineering perspective*. Chapman and Hall/CRC.
- Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, 156, 134-142.
<https://www.doi.org/10.1016/j.neucom.2014.12.073>
- Ghorbani, S., & Desmarais, M. C. (2017). Performance comparison of recent imputation methods for classification tasks over binary data. *Applied Artificial Intelligence*, 31(1), 1-22. <https://www.doi.org/10.1080/08839514.2017.1279046>
- González-Ladrón-de-Guevara, F., Fernández-Diego, M., & Lokan, C. (2016). The usage of ISBSG data fields in software effort estimation: A systematic mapping study. *Journal of Systems and Software*, 113, 188-215.
<https://www.doi.org/10.1016/j.jss.2015.11.040>
- Government Accountability Office. (1972). *Theory and Practice of Cost Estimating for Major Acquisitions* (Report No. B-163508). U.S. Government Printing Office.
<http://www.gao.gov/assets/210/200036.pdf>
- Government Accountability Office. (2009). *GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs* (Report No. GAO-09-3SP). U.S. Government Printing Office.
<https://www.gao.gov/new.items/d093sp.pdf>
- Government Accountability Office. (2020). *GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Program Costs* (Report No.

GAO-20-195G). U.S. Government Printing Office.

<https://www.gao.gov/assets/710/705312.pdf>

Govinfo. (2020). *A budget for a better America: Promises kept, taxpayers first*. U.S.

Government Publishing Office. <https://www.govinfo.gov/content/pkg/BUDGET-2020-BUD/pdf/BUDGET-2020-BUD.pdf>

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation

theory. *Prevention science*, 8(3), 206-213. <https://www.doi.org/10.1007/s11121-007-0070-9>

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576.

<https://www.doi.org/10.1146/annurev.psych.58.110405.085530>

Graham, J. W. (2012). Missing data theory. In *Missing Data* (pp. 3-46). Springer.

https://doi.org/10.1007/978-1-4614-4018-5_1

Grimstad, S., Jorgensen, M., & Molokken-Ostfold, K. (2006). Software effort estimation terminology: The tower of Babel. *Information and Software Technology*, 48(4),

302-310. <https://www.doi.org/10.1016/j.infsof.2005.04.004>

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79-90. <https://doi.org/10.1198/000313007X172556>

- Huang, J., Li, Y. F., Keung, J. W., Yu, Y. T., & Chan, W. K. (2017, July). An empirical analysis of three-stage data-preprocessing for analogy-based software effort estimation on the ISBSG data. In *Software Quality, Reliability and Security (QRS), 2017 IEEE International Conference on* (pp. 442-449). IEEE.
<https://www.doi.org/10.1109/QRS.2017.54>
- Huang, J., Li, Y. F., & Xie, M. (2015a). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67, 108-127. <https://www.doi.org/10.1016/j.infsof.2015.07.004>
- Huang, J., Sun, H., Li, Y. F., & Xie, M. (2015b, August). An empirical study of dynamic incomplete-case nearest neighbor imputation in software quality data. In *2015 IEEE International Conference on Software Quality, Reliability and Security* (pp. 37-42). IEEE. <https://www.doi.org/10.1109/QRS.2015.16>
- IBM knowledge center. (2021). IBM - United States.
https://www.ibm.com/support/knowledgecenter/SSLVMB_sub/statistics_mainhelp_ddita/spss/mva/idh_miss_em.html
- Idri, A., Abnane, I., & Abran, A. (2015a, June). Systematic mapping study of missing values techniques in software engineering data. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on* (pp. 1-8). IEEE.
<https://www.doi.org/10.1109/SNPD.2015.7176280>
- Idri, A., Abnane, I., & Abran, A. (2016a). Dealing with missing values in software project datasets: A systematic mapping study. In *Software Engineering, Artificial*

Intelligence, Networking and Parallel/Distributed Computing (pp. 1-16). Springer

https://www.doi.org/10.1007/978-3-319-33810-1_1

Idri, A., Abnane, I., & Abran, A. (2016b). Missing data techniques in analogy-based

software development effort estimation. *The Journal of Systems & Software*, 117,

595-611. <https://www.doi.org/10.1016/j.jss.2016.04.058>

Idri, A., azzahra Amazal, F., & Abran, A. (2015b). Analogy-based software development

effort estimation: A systematic mapping and review. *Information and Software*

Technology, 58, 206-230. <https://www.doi.org/10.1016/j.infsof.2014.07.013>

Idri, A., azzahra Amazal, F., & Abran, A. (2016c). Accuracy comparison of analogy-

based software development effort estimation techniques. *International Journal of*

Intelligent Systems, 31(2), 128-152. <https://www.doi.org/10.1002/int.21748>

Iqbal, S. Z., Idrees, M., Sana, A. B., & Khan, N. (2017). Comparative analysis of

common software cost estimation modelling techniques. *Mathematical Modelling*

and Applications, 2(3), 33. <https://www.doi.org/10.11648/j.mma.20170203.12>

International Cost Estimation and Analysis Association. (2019). *International Cost*

Estimation and Analysis Association Testable Topics List.

<http://www.iceaaonline.com/ready/wpcontent/uploads/2014/02/testableTopicsList.pdf>

IT Dashboarddata.gov (2021). *Archived 2013-2017 IT agency data sets*.

<https://itdashboard.gov/drupal/archived-data>

- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10), 913-933. <https://www.doi.org/10.1080/08839514.2019.1637138>
- Janssen, K. J., Donders, A. R. T., Harrell Jr., F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, 63(7), 721-727. <https://www.doi.org/10.1016/j.jclinepi.2009.12.008>
- Jeffery, R., Ruhe, M., & Wieczorek, I. (2000). A comparative study of cost modelling techniques using public domain multi-organizational and company-specific data. In *Proc. of the European Software Control and Metrics Conference (ESCOM)* (Vol. 2000). https://www.researchgate.net/profile/I_Wieczorek/publication/242385779_A_comparative_Study_of_Cost_Modelling_Techniques_using_Public_Domain_multi-organisational_and_company-specific_Data/links/0a85e52d6af2e756c4000000.pdf
- Jeffery, R., Ruhe, M., & Wieczorek, I. (2001, April). Using public domain metrics to estimate software development effort. In *Proceedings Seventh International Software Metrics Symposium* (pp. 16-27). IEEE. <https://www.doi.org/10.1109/METRIC.2001.915512>
- Jing, X. Y., Qi, F., Wu, F., and Xiu, B. (2016). Missing data imputation based on low-rank recovery and semi-supervised regression for software effort estimation. *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*,

Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on, ICSE, 607. <https://www.doi.org/10.1145/2884781.2884827>

Joint Agency Cost Estimating Relationship (CER) Development Handbook. (2018, February 9).

<https://www.asafm.army.mil/Portals/72/Documents/Offices/CE/CER%20Development%20Handbook.pdf>

Jones, C. (2007). *Estimating software costs: Bringing realism to estimating*. McGraw-Hill Companies.

Jorgensen, M. (2006). How large are software cost overruns? A review of the 1994 CHAOS report. *Information and Software Technology*, 48(4), 297-301.

<https://www.doi.org/10.1016/j.infsof.2005.07.002>

Kapelner, A., & Bleich, J. (2015). Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2), 224-239.

<https://www.doi.org/10.1002/cjs.11248>

Kendall, F. (2013). Implementation directive for better buying power 3.0 – Achieving greater efficiency and productivity in defense spending

https://business.defense.gov/Portals/57/Documents/Attachment%209_BBP%20%20Implementation%20Directive.pdf

Kendall, F. (2015). Implementation directive for better buying power 3.0 – Achieving dominant capabilities through technical excellence and innovation.

[https://www.acq.osd.mil/fo/docs/betterBuyingPower3.0\(9Apr15\).pdf](https://www.acq.osd.mil/fo/docs/betterBuyingPower3.0(9Apr15).pdf)

- Khoshgoftaar, T. M., & Van Hulse, J. (2008). Imputation techniques for multivariate missingness in software measurement data. *Software Quality Journal*, 16(4), 563-600. <https://www.doi.org/10.1007/s11219-008-9054-7>
- Kiasari, M. A., Jang, G. J., & Lee, M. (2017). Novel iterative approach using generative and discriminative models for classification with missing features. *Neurocomputing*, 225, 23-30.
<https://www.doi.org/10.1016/j.neucom.2016.11.015>
- Kirk, R. E. (2013). Chapter 5 Multiple Comparison Tests. In R.E. Kirk (Ed.), *Experimental design: Procedures for the behavioral sciences* (pp. 154-208) Sage Publications. <https://www.doi.org/10.4135/9781483384733>
- Kreinovich, V. (2012). How to define relative approximation error of an interval estimate: a proposal.
<http://www.cs.utep.edu/vladik/2012/tr12-37.pdf>
- Li, Y., & Parker, L. E. (2014). Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks. *Information Fusion*, 15, 64-79.
<https://www.doi.org/10.1016/j.inffus.2012.08.007>
- Li, Z., Sharaf, M. A., Sitbon, L., Sadiq, S., Indulska, M., & Zhou, X. (2014). A web-based approach to data imputation. *World Wide Web*, 17(5), 873-897.
<https://www.doi.org/10.1007/s11280-013-0263-z>
- Lin, W. C., & Tsai, C. F. (2019). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 1-23.
<https://www.doi.org/10.1007/s10462-019-09709-4>

- Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Vol. 2). John Wiley & Sons.
- Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data* (Vol. 3). John Wiley & Sons.
- MacDonell, S. G., & Shepperd, M. J. (2003). Combining techniques to optimize effort predictions in software project management. *Journal of Systems and Software*, 66(2), 91-98. [https://www.doi.org/10.1016/S0164-1212\(02\)00067-5](https://www.doi.org/10.1016/S0164-1212(02)00067-5)
- Majeed, F. (2018, February). Model based estimation approach to the missing data problem. In *Advances in Science and Engineering Technology International Conferences (ASET), 2018* (pp. 1-6). IEEE. <https://www.doi.org/10.1109/ICASET.2018.8376847>
- Melese, F., Richter, A., & Solomon, B. (Eds.). (2015). *Military cost–benefit analysis: Theory and practice*. Routledge.
- Mesquite, D. P., Gomes, J. P., Junior, A. H. S., & Nobre, J. S. (2017). Euclidean distance estimation in incomplete datasets. *Neurocomputing*, 248 (2017), 11-18. <https://www.doi.org/10.1016/j.neucom.2016.12.081>
- Mislick, G. K., & Nussbaum, D. A. (2015). *Cost estimation: Methods and tools*. John Wiley & Sons.
- Mittas, N., & Angelis, L. (2008, October). Combining regression and estimation by

analogy in a semi-parametric model for software cost estimation. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement* (pp. 70-79). ACM.

<https://www.doi.org/10.1145/1414004.1414017>

Morin, J. M. (2017, January 9). DOD cost analysis data improvement.

<http://www.acqnotes.com/wp-content/uploads/2014/09/DoD-Cost-Analysis-Data-Improvement-Memo-Signed-by-Dr-Morin-2017-01-09-002.pdf>

Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*, 27(11), pp. 999-1013.

<https://www.doi.org/10.1109/32.965340>

Myrtveit, I., Stensrud, E., & Olsson, U. (2001). Assessing the benefits of imputing ERP projects with missing data. In *Proceedings Seventh International Software Metrics Symposium* (pp. 78-84). IEEE.

<https://www.doi.org/10.1109/METRIC.2001.915517>

Nagashima, H., & Kato, Y. (2019, March). APREP-DM: A Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (pp. 555-560). IEEE.

<https://www.doi.org/10.1109/PERCOMW.2019.8730785>

Newnan, D. G., Eschenbach, T., & Lavelle, J. P. (2004). *Study Guide for Engineering Economic Analysis*. Oxford University Press.

- Office of the Under Secretary of Defense (OUSD) for Acquisition, Technology and Logistics (AT&L). (2019). Acquisition resources and analysis (ARA) directorate MDAP and MAIS list.
https://www.acq.osd.mil/ara/documents/mdap_mais_program_list.pdf
- Ostwald, P. F. (1974). *Cost estimating for engineering and management*. Prentice-Hall.
- Parnell, G. S. (2017). *Trade-off analytics: creating and exploring the system trade space*. John Wiley & Sons.
- Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, 42(13), 5621-5631. <https://www.doi.org/10.1016/j.eswa.2015.02.050>
- Qi, F., Jing, X. Y., Zhu, X., Xie, X., Xu, B., & Ying, S. (2017). Software effort estimation based on open-source projects: Case study of Github. *Information and Software Technology*, 92, 145-157.
<https://www.doi.org/10.1016/j.infsof.2017.07.015>
- Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2009). POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert systems with applications*, 36(2), 2794-2804.
- Research Randomizer. (2020). <https://randomizer.org>
- Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford Publications.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
<https://www.doi.org/10.1093/biomet/63.3.581>

- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 1). John Wiley & Sons.
- Saeed, A., Butt, W. H., Kazmi, F., & Arif, M. (2018, February). Survey of software development effort estimation techniques. In *Proceedings of the 2018 7th International Conference on Software and Computer Applications* (pp. 82-86). ACM. <https://www.doi.org/10.1145/3185089.3185140>
- Sentas, P., & Angelis, L. (2006). Categorical missing data imputation for software cost estimation by multinomial logistic regression. *The Journal of Systems & Software*, 79 (3), 404-414. <https://www.doi.org/10.1016/j.jss.2005.02.026>
- Seo, Y. S., Yoon, K. A., & Bae, D. H. (2009, December). Improving the accuracy of software effort estimation based on multiple least square regression models by estimation error-based data partitioning. In *2009 16th Asia-Pacific Software Engineering Conference* (pp. 3-10). IEEE. <https://www.doi.org/10.1109/APSEC.2009.57>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7(2), 147-177. <https://www.doi.org/10.1037//1082989X.7.2.147>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shek D. T. & Zhu, X. (2018). Pretest-posttest designs. In *The SAGE encyclopedia of*

educational research, measurement, and evaluation. (pp. 1293-1295). SAGE

Publications. <https://www.doi.org/10.4135/9781506326139>

Schwartz, M., & O'Connor, C. V. (2016). The Nunn-McCurdy Act: Background, analysis, and issues for Congress (Report No. R41293). Congressional Research Service.

https://www.everycrsreport.com/files/20160512_R41293_98924a507e8ec5a0bfe6dac75c6c5dac3ad38b1f.pdf

Singleton, R. A., & Straits, B. C. 2010. *Approaches to Social Research*.

Oxford University Press.

Spruill, N. (2021). Business—Cost Estimating and financial management functional communities. <https://www.dau.edu/training/career-development/bce/p/Business-Cost-Estimating-and-Financial-Management-Functional-Communities>

Soltanveis, F., & Alizadeh, S. H. (2016, April). Using parametric regression and KNN algorithm with missing handling for software effort prediction. In *2016 Artificial Intelligence and Robotics (IRANOPEN)* (pp. 77-84). IEEE.

<https://www.doi.org/10.1109/RIOS.2016.7529494>

Song, Q., Shepperd, M., Chen, X., & Liu, J. (2008). Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. *Journal of Systems and Software*, *81*(12), 2361-2370.

<https://www.doi.org/10.1016/j.jss.2008.05.008>

Strike, K., El Emam, K., & Madhavji, N. (2001). Software cost estimation with

incomplete data. *IEEE Transactions on Software Engineering*, 27(10), 890-908.

<https://www.doi.org/10.1109/32.962560>

Thyer, B. A. (2012). *Quasi-experimental research designs*. Oxford University Press.

Trochim, W. M., & Donnelly, J. P. (2008). The research methods knowledge base.

Atomic Dog/Cengage Learning.

Trochim, W. M., Donnelly, J. P., & Arora, K. (2016). The essential research methods knowledge base. *Cengage Learning*.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of operations management*, 24(1), 53-62.

<https://www.doi.org/10.1016/j.jom.2005.03.001>

Twala, B. (2017). When partly missing data matters in software effort development prediction. *Journal of advanced computational intelligence and intelligent informatics*, 21(5), 803-812.

<https://www.doi.org/10.20965/jaciii.2017.p0803>

Twala, B., Cartwright, M., & Shepperd, M. (2006, May). Ensemble of missing data techniques to improve software prediction accuracy. In Proceedings of the 28th international conference on Software engineering (pp. 909-912).

<https://www.doi.org/10.1145/1134285.1134449>

10 U.S. Code § 1746 – Defense Acquisition University. (2012).

<https://www.govinfo.gov/content/pkg/USCODE-2011-title10/html/USCODE-2011-title10-subtitleA-partII-chap87-subchapIV-sec1746.htm>

10 U.S. Code § 2334 – Independent cost estimation and cost analysis. (2017).

<https://www.govinfo.gov/app/details/USCODE-2017-title10/USCODE-2017-title10-subtitleA-partIV-chap137-sec2334>

17 U.S. Code § 105 – Subject matter of copyright: United States Government works.

(2010). <https://www.govinfo.gov/app/details/USCODE-2010-title17/USCODE-2010-title17-chap1-sec105>

Undersecretary of Defense for Acquisition and Sustainment (2019). DoD Major Defense Acquisition Program and Major Automated Information Systems (MAIS) List.

https://www.acq.osd.mil/ara/documents/mdap_mais_program_list.pdf

University of California at San Francisco (UCSF). (2019). Sample size calculators:

Effect size for before-after-study (Paired T-test). <http://www.sample-size.net/effect-size-study-paired-t-test/>

USASpending.gov. (2021). <https://www.usaspending.gov/#/>

Valerdi, R., Dabkowski, M., & Dixit, I. (2015). Reliability improvement of major defense acquisition program cost estimates—Mapping DoDAF to COSYSMO. *Systems Engineering*, 18(5), 530-547. <https://www.doi.org/10.1002/sys.21327>

Van Hulse, J., & Khoshgoftaar, T. M. (2014). Incomplete-case nearest neighbor imputation in software measurement data. *Information Sciences*, 259, 596-610. <https://www.doi.org/10.1016/j.ins.2010.12.017>

Walden University. (2020). Form A: First step of ethics review (2019).

<https://www.emailmeform.com/builder/form/Kel0pm0Cf9e2D4GAI18e>

Wani, Z. H., Giri, K. J., & Bashir, R. (2019). A generic data mining model for software cost estimation based on novel input selection procedure. *International Journal of*

Information Retrieval Research (IJIRR), 9(1), 16-32

Williams, T., & Barber, E. (2011). *Cost Estimation Methodologies*.

<https://www.dau.mil/cop/ce/DAU%20Sponsored%20Documents/B4%20CE%20Methodologies%20Feb%202011%20V3.pdf>

Zhang, W., Yang, Y., & Wang, Q. (2011, September). Handling missing data in software effort prediction with naive Bayes and EM algorithm. In *7th International Conference on Predictive Models in Software Engineering, PROMISE 2011, Co-located with ESEM 2011* (1-10). <https://www.doi.org/10.1145/2020390.2020394>

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2010). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, 23(1), 110-121. <https://www.doi.org/10.1109/TKDE.2010.99>

Appendix A: Closest Predictive Accuracy Results Per Data Set

Appendix A presents the results of each pre-experimental run per data set. Tables A1-A28 list the individual results per each “Experiment Run Case No.” within a data set and notate which missing data theory techniques calculated’ “Predicted Value” came in closest to the “Original Numerical Value”. The “1” in each column denotes which technique had the lowest absolute error and relative error compared to each individual unique synthetic software program’s data value that was removed-at-random and replaced with a new value based on the applied missing data theory technique’s “Predicted Value”. Summary totals are captured in the headers on each table for listwise delete (LD), single imputation using the mean (SI-Mean), and multiple imputation using linear regression (MI-LR).

Table A1

Closest Missing Data Theory Predictive Accuracy in Data Set 1 (DS1) Experiments for Number of External Interfaces Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (10)	MI-LR (46)
1	0.05P31			1
2	0.05P91		1	
3	0.10P21			1
4	0.10P91		1	
5	0.10P181			1
6	0.15P31			1
7	0.15P61			1
8	0.15P71			1
9	0.15P241			1
10	0.15P301			1
11	0.20P21			1
12	0.20P31			1
13	0.20P91		1	
14	0.20P121		1	

15	0.20P181		1
16	0.20P291		1
17	0.25P61		1
18	0.25P81	1	
19	0.25P111	1	
20	0.25P121	1	
21	0.25P141	1	
22	0.25P171	1	
23	0.25P251		1
24	0.25P271		1
25	0.30P11		1
26	0.30P21		1
27	0.30P161		1
28	0.30P181		1
29	0.30P191		1
30	0.30P211		1
31	0.30P241		1
32	0.30P251		1
33	0.30P301		1
34	0.35P11		1
35	0.35P21		1
36	0.35P31		1
37	0.35P71		1
38	0.35P81		1
39	0.35P151	1	
40	0.35P181		1
41	0.35P191		1
42	0.35P251		1
43	0.35P281		1
44	0.35P291		1
45	0.40P11		1
46	0.40P21		1
47	0.40P31		1
48	0.40P151		1
49	0.40P161		1
50	0.40P171		1
51	0.40P181		1
52	0.40P191		1
53	0.40P201		1
54	0.40P251		1
55	0.40P291		1
56	0.40P301		1

Table A2

Closest Missing Data Theory Predictive Accuracy in Data Set 2 (DS2) Experiments for Initial Software Lines of Code (SLOC)-New Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (19)	MI-LR (37)
1	0.05P32			1
2	0.05P92			1
3	0.10P22			1
4	0.10P92			1
5	0.10P182		1	
6	0.15P32			1
7	0.15P62			1
8	0.15P72		1	
9	0.15P242			1
10	0.15P302		1	
11	0.20P22			1
12	0.20P32		1	
13	0.20P92			1
14	0.20P122		1	
15	0.20P182		1	
16	0.20P292			1
17	0.25P62			1
18	0.25P82		1	
19	0.25P112			1
20	0.25P122		1	
21	0.25P142		1	
22	0.25P172		1	
23	0.25P252			1
24	0.25P272			1
25	0.30P12			1
26	0.30P22			1
27	0.30P162			1
28	0.30P182		1	
29	0.30P192			1
30	0.30P212		1	
31	0.30P242			1
32	0.30P252			1
33	0.30P302		1	
34	0.35P12			1
35	0.35P22			1

36	0.35P32		1
37	0.35P72	1	
38	0.35P82	1	
39	0.35P152		1
40	0.35P182	1	
41	0.35P192		1
42	0.35P252		1
43	0.35P282		1
44	0.35P292		1
45	0.40P12		1
46	0.40P22		1
47	0.40P32		1
48	0.40P152		1
49	0.40P162		1
50	0.40P172		1
51	0.40P182	1	
52	0.40P192		1
53	0.40P202	1	
54	0.40P252		1
55	0.40P292		1
56	0.40P302	1	

Table A3

Closest Missing Data Theory Predictive Accuracy in Data Set 3 (DS3) Experiments for Initial SLOC Modified Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (31)	MI-LR (25)
1	0.05P33		1	
2	0.05P93			1
3	0.1P23		1	
4	0.1P93		1	
5	0.1P183		1	
6	0.15P33		1	
7	0.15P63			1
8	0.15P73		1	
9	0.15P243		1	
10	0.15P303			1
11	0.2P23			1
12	0.2P33			1
13	0.2P93		1	

14	0.2P123		1
15	0.2P183	1	
16	0.2P293		1
17	0.25P63		1
18	0.25P83	1	
19	0.25P113	1	
20	0.25P123	1	
21	0.25P143	1	
22	0.25P173	1	
23	0.25P253		1
24	0.25P273	1	
25	0.3P13	1	
26	0.3P23	1	
27	0.3P163		1
28	0.3P183		1
29	0.3P193	1	
30	0.3P213	1	
31	0.3P243	1	
32	0.3P253	1	
33	0.3P303		1
34	0.35P13		1
35	0.35P23		1
36	0.35P33		1
37	0.35P73	1	
38	0.35P83		1
39	0.35P153	1	
40	0.35P183	1	
41	0.35P193	1	
42	0.35P253		1
43	0.35P283	1	
44	0.35P293		1
45	0.4P13		1
46	0.4P23		1
47	0.4P33		1
48	0.4P153	1	
49	0.4P163		1
50	0.4P173	1	
51	0.4P183	1	
52	0.4P193	1	
53	0.4P203	1	
54	0.4P253		1
55	0.4P293		1
56	0.4P303		1

Table A4

Closest Missing Data Theory Predictive Accuracy in Data Set 4 (DS4) Experiments for Initial SLOC Reused Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (30)	MI-LR (26)
1	0.05P34		1	
2	0.05P94		1	
3	0.1P24			1
4	0.1P94			1
5	0.1P184		1	
6	0.15P34			1
7	0.15P64			1
8	0.15P74			1
9	0.15P244		1	
10	0.15P304			1
11	0.2P24		1	
12	0.2P34			1
13	0.2P94		1	
14	0.2P124		1	
15	0.2P184		1	
16	0.2P294			1
17	0.25P64			1
18	0.25P84			1
19	0.25P114		1	
20	0.25P124		1	
21	0.25P144		1	
22	0.25P174		1	
23	0.25P254		1	
24	0.25P274		1	
25	0.3P14			1
26	0.3P24			1
27	0.3P164		1	
28	0.3P184		1	
29	0.3P194		1	
30	0.3P214			1
31	0.3P244		1	
32	0.3P254		1	
33	0.3P304			1
34	0.35P14		1	
35	0.35P24			1
36	0.35P34			1
37	0.35P74			1

38	0.35P84	1	
39	0.35P154	1	
40	0.35P184	1	
41	0.35P194	1	
42	0.35P254	1	
43	0.35P284		1
44	0.35P294		1
45	0.4P14		1
46	0.4P24		1
47	0.4P34		1
48	0.4P154	1	
49	0.4P164	1	
50	0.4P174	1	
51	0.4P184	1	
52	0.4P194		1
53	0.4P204		1
54	0.4P254	1	
55	0.4P294		1
56	0.4P304		1

Table A5

Closest Missing Data Theory Predictive Accuracy in Data Set 5 (DS5) Experiments for Final SLOC – New Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (25)	MI-LR (31)
1	0.05P35		1	
2	0.05P95			1
3	0.1P25			1
4	0.1P95		1	
5	0.1P185		1	
6	0.15P35			1
7	0.15P65			1
8	0.15P75			1
9	0.15P245			1
10	0.15P305		1	
11	0.2P25		1	
12	0.2P35		1	
13	0.2P95			1
14	0.2P125			1
15	0.2P185		1	

16	0.2P295		1
17	0.25P65		1
18	0.25P85	1	
19	0.25P115	1	
20	0.25P125	1	
21	0.25P145	1	
22	0.25P175		1
23	0.25P255		1
24	0.25P275	1	
25	0.3P15		1
26	0.3P25		1
27	0.3P165	1	
28	0.3P185		1
29	0.3P195		1
30	0.3P215	1	
31	0.3P245	1	
32	0.3P255		1
33	0.3P305	1	
34	0.35P15		1
35	0.35P25	1	
36	0.35P35	1	
37	0.35P75		1
38	0.35P85		1
39	0.35P155	1	
40	0.35P185		1
41	0.35P195		1
42	0.35P255		1
43	0.35P285		1
44	0.35P295		1
45	0.4P15		1
46	0.4P25	1	
47	0.4P35		1
48	0.4P155		1
49	0.4P165		1
50	0.4P175	1	
51	0.4P185	1	
52	0.4P195		1
53	0.4P205	1	
54	0.4P255	1	
55	0.4P295		1
56	0.4P305	1	

Table A6

Closest Missing Data Theory Predictive Accuracy in Data Set 6 (DS6) Experiments for Final SLOC – Modified Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (29)	MI-LR (27)
1	0.05P36		1	
2	0.05P96			1
3	0.1P26			1
4	0.1P96		1	
5	0.1P186		1	
6	0.15P36		1	
7	0.15P66		1	
8	0.15P76		1	
9	0.15P246		1	
10	0.15P306			1
11	0.2P26			1
12	0.2P36			1
13	0.2P96			1
14	0.2P126			1
15	0.2P186		1	
16	0.2P296			1
17	0.25P66			1
18	0.25P86		1	
19	0.25P116		1	
20	0.25P126		1	
21	0.25P146		1	
22	0.25P176			1
23	0.25P256			1
24	0.25P276			1
25	0.3P16			1
26	0.3P26			1
27	0.3P166			1
28	0.3P186		1	
29	0.3P196		1	
30	0.3P216		1	
31	0.3P246		1	
32	0.3P256		1	
33	0.3P306			1
34	0.35P16			1
35	0.35P26			1

36	0.35P36	1	
37	0.35P76		1
38	0.35P86		1
39	0.35P156		1
40	0.35P186	1	
41	0.35P196		1
42	0.35P256		1
43	0.35P286	1	
44	0.35P296		1
45	0.4P16	1	
46	0.4P26	1	
47	0.4P36	1	
48	0.4P156	1	
49	0.4P166		1
50	0.4P176	1	
51	0.4P186	1	
52	0.4P196	1	
53	0.4P206	1	
54	0.4P256	1	
55	0.4P296		1
56	0.4P306		1

Table A7

Closest Missing Data Theory Predictive Accuracy in Data Set 7 (DS7) Experiments for Final SLOC – Reused Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (34)	MI-LR (22)
1	0.05P37			1
2	0.05P97			1
3	0.1P27			1
4	0.1P97		1	
5	0.1P187		1	
6	0.15P37			1
7	0.15P67			1
8	0.15P77			1
9	0.15P247		1	
10	0.15P307			1
11	0.2P27		1	
12	0.2P37		1	
13	0.2P97		1	

14	0.2P127		1
15	0.2P187	1	
16	0.2P297		1
17	0.25P67	1	
18	0.25P87	1	
19	0.25P117	1	
20	0.25P127		1
21	0.25P147	1	
22	0.25P177	1	
23	0.25P257	1	
24	0.25P277	1	
25	0.3P17	1	
26	0.3P27		1
27	0.3P167		1
28	0.3P187		1
29	0.3P197	1	
30	0.3P217		1
31	0.3P247	1	
32	0.3P257	1	
33	0.3P307		1
34	0.35P17	1	
35	0.35P27	1	
36	0.35P37	1	
37	0.35P77		1
38	0.35P87	1	
39	0.35P157	1	
40	0.35P187		1
41	0.35P197	1	
42	0.35P257	1	
43	0.35P287		1
44	0.35P297		1
45	0.4P17	1	
46	0.4P27		1
47	0.4P37	1	
48	0.4P157	1	
49	0.4P167	1	
50	0.4P177	1	
51	0.4P187	1	
52	0.4P197	1	
53	0.4P207	1	
54	0.4P257	1	
55	0.4P297		1
56	0.4P307		1

Table A8

Closest Missing Data Theory Predictive Accuracy in Data Set 8 (DS8) Experiments for Re-Design/ Design Modified Effort (DM) % - Modified Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (27)	MI-LR (29)
1	0.05P38		1	
2	0.05P98			1
3	0.1P28		1	
4	0.1P98			1
5	0.1P188			1
6	0.15P38			1
7	0.15P68			1
8	0.15P78			1
9	0.15P248			1
10	0.15P308			1
11	0.2P28			1
12	0.2P38		1	
13	0.2P98			1
14	0.2P128			1
15	0.2P188		1	
16	0.2P298		1	
17	0.25P68		1	
18	0.25P88		1	
19	0.25P118		1	
20	0.25P128			1
21	0.25P148		1	
22	0.25P178			1
23	0.25P258			1
24	0.25P278			1
25	0.3P18		1	
26	0.3P28		1	
27	0.3P168		1	
28	0.3P188		1	
29	0.3P198			1
30	0.3P218		1	
31	0.3P248		1	
32	0.3P258		1	
33	0.3P308			1
34	0.35P18			1
35	0.35P28		1	
36	0.35P38			1
37	0.35P78		1	

38	0.35P88		1
39	0.35P158	1	
40	0.35P188	1	
41	0.35P198	1	
42	0.35P258	1	
43	0.35P288		1
44	0.35P298	1	
45	0.4P18	1	
46	0.4P28		1
47	0.4P38		1
48	0.4P158	1	
49	0.4P168		1
50	0.4P178		1
51	0.4P188		1
52	0.4P198		1
53	0.4P208		1
54	0.4P258	1	
55	0.4P298		1
56	0.4P308	1	

Table A9

Closest Missing Data Theory Predictive Accuracy in Data Set 9 (DS9) Experiments for Re-Code/ Code Modified Effort (CM) % - Modified Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (31)	MI-LR (25)
1	0.05P39			1
2	0.05P99		1	
3	0.1P29		1	
4	0.1P99		1	
5	0.1P189		1	
6	0.15P39		1	
7	0.15P69		1	
8	0.15P79		1	
9	0.15P249			1
10	0.15P309		1	
11	0.2P29			1
12	0.2P39			1
13	0.2P99		1	
14	0.2P129			1
15	0.2P189		1	

16	0.2P299	1	
17	0.25P69	1	
18	0.25P89		1
19	0.25P119	1	
20	0.25P129	1	
21	0.25P149	1	
22	0.25P179	1	
23	0.25P259	1	
24	0.25P279	1	
25	0.3P19		1
26	0.3P29	1	
27	0.3P169	1	
28	0.3P189		1
29	0.3P199		1
30	0.3P219		1
31	0.3P249	1	
32	0.3P259		1
33	0.3P309	1	
34	0.35P19		1
35	0.35P29	1	
36	0.35P39		1
37	0.35P79	1	
38	0.35P89		1
39	0.35P159		1
40	0.35P189		1
41	0.35P199	1	
42	0.35P259	1	
43	0.35P289	1	
44	0.35P299	1	
45	0.4P19		1
46	0.4P29	1	
47	0.4P39		1
48	0.4P159		1
49	0.4P169		1
50	0.4P179		1
51	0.4P189		1
52	0.4P199		1
53	0.4P209		1
54	0.4P259	1	
55	0.4P299	1	
56	0.4P309		1

Table A10

Closest Missing Data Theory Predictive Accuracy in Data Set 10 (DS10) Experiments for Re-Test/ Integration Modified Effort (IM) % - Modified Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (31)	MI-LR (25)
1	0.05P310			1
2	0.05P910			1
3	0.1P210		1	
4	0.1P910			1
5	0.1P1810			1
6	0.15P310		1	
7	0.15P610			1
8	0.15P710		1	
9	0.15P2410		1	
10	0.15P3010		1	
11	0.2P210		1	
12	0.2P310		1	
13	0.2P910		1	
14	0.2P1210		1	
15	0.2P1810		1	
16	0.2P2910			1
17	0.25P610			1
18	0.25P810			1
19	0.25P1110			1
20	0.25P1210		1	
21	0.25P1410			1
22	0.25P1710		1	
23	0.25P2510			1
24	0.25P2710		1	
25	0.3P110			1
26	0.3P210		1	
27	0.3P1610		1	
28	0.3P1810		1	
29	0.3P1910		1	
30	0.3P2110			1
31	0.3P2410			1
32	0.3P2510		1	
33	0.3P3010			1
34	0.35P110			1
35	0.35P210		1	
36	0.35P310		1	
37	0.35P710			1

38	0.35P810	1	
39	0.35P1510	1	
40	0.35P1810	1	
41	0.35P1910		1
42	0.35P2510		1
43	0.35P2810	1	
44	0.35P2910	1	
45	0.4P110		1
46	0.4P210	1	
47	0.4P310	1	
48	0.4P1510		1
49	0.4P1610	1	
50	0.4P1710	1	
51	0.4P1810	1	
52	0.4P1910		1
53	0.4P2010	1	
54	0.4P2510		1
55	0.4P2910		1
56	0.4P3010		1

Table A11

Closest Missing Data Theory Predictive Accuracy in Data Set 11 (DS11) Experiments for Design Modified (DM) % - Reused Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (56)	MI-LR (56)
1	0.05P311		Perfect	Perfect
2	0.05P911		Perfect	Perfect
3	0.1P211		Perfect	Perfect
4	0.1P911		Perfect	Perfect
5	0.1P1811		Perfect	Perfect
6	0.15P311		Perfect	Perfect
7	0.15P611		Perfect	Perfect
8	0.15P711		Perfect	Perfect
9	0.15P2411		Perfect	Perfect
10	0.15P3011		Perfect	Perfect
11	0.2P211		Perfect	Perfect
12	0.2P311		Perfect	Perfect
13	0.2P911		Perfect	Perfect
14	0.2P1211		Perfect	Perfect
15	0.2P1811		Perfect	Perfect
16	0.2P2911		Perfect	Perfect

17	0.25P611	Perfect	Perfect
18	0.25P811	Perfect	Perfect
19	0.25P1111	Perfect	Perfect
20	0.25P1211	Perfect	Perfect
21	0.25P1411	Perfect	Perfect
22	0.25P1711	Perfect	Perfect
23	0.25P2511	Perfect	Perfect
24	0.25P2711	Perfect	Perfect
25	0.3P111	Perfect	Perfect
26	0.3P211	Perfect	Perfect
27	0.3P1611	Perfect	Perfect
28	0.3P1811	Perfect	Perfect
29	0.3P1911	Perfect	Perfect
30	0.3P2111	Perfect	Perfect
31	0.3P2411	Perfect	Perfect
32	0.3P2511	Perfect	Perfect
33	0.3P3011	Perfect	Perfect
34	0.35P111	Perfect	Perfect
35	0.35P211	Perfect	Perfect
36	0.35P311	Perfect	Perfect
37	0.35P711	Perfect	Perfect
38	0.35P811	Perfect	Perfect
39	0.35P1511	Perfect	Perfect
40	0.35P1811	Perfect	Perfect
41	0.35P1911	Perfect	Perfect
42	0.35P2511	Perfect	Perfect
43	0.35P2811	Perfect	Perfect
44	0.35P2911	Perfect	Perfect
45	0.4P111	Perfect	Perfect
46	0.4P211	Perfect	Perfect
47	0.4P311	Perfect	Perfect
48	0.4P1511	Perfect	Perfect
49	0.4P1611	Perfect	Perfect
50	0.4P1711	Perfect	Perfect
51	0.4P1811	Perfect	Perfect
52	0.4P1911	Perfect	Perfect
53	0.4P2011	Perfect	Perfect
54	0.4P2511	Perfect	Perfect
55	0.4P2911	Perfect	Perfect
56	0.4P3011	Perfect	Perfect

Table A12

Closest Missing Data Theory Predictive Accuracy in Data Set 12 (DS12) Experiments for Code Modified (CM) % - Reused Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (56)	MI-LR (56)
1	0.05P312		Perfect	Perfect
2	0.05P912		Perfect	Perfect
3	0.1P212		Perfect	Perfect
4	0.1P912		Perfect	Perfect
5	0.1P1812		Perfect	Perfect
6	0.15P312		Perfect	Perfect
7	0.15P612		Perfect	Perfect
8	0.15P712		Perfect	Perfect
9	0.15P2412		Perfect	Perfect
10	0.15P3012		Perfect	Perfect
11	0.2P212		Perfect	Perfect
12	0.2P312		Perfect	Perfect
13	0.2P912		Perfect	Perfect
14	0.2P1212		Perfect	Perfect
15	0.2P1812		Perfect	Perfect
16	0.2P2912		Perfect	Perfect
17	0.25P612		Perfect	Perfect
18	0.25P812		Perfect	Perfect
19	0.25P1112		Perfect	Perfect
20	0.25P1212		Perfect	Perfect
21	0.25P1412		Perfect	Perfect
22	0.25P1712		Perfect	Perfect
23	0.25P2512		Perfect	Perfect
24	0.25P2712		Perfect	Perfect
25	0.3P112		Perfect	Perfect
26	0.3P212		Perfect	Perfect
27	0.3P1612		Perfect	Perfect
28	0.3P1812		Perfect	Perfect
29	0.3P1912		Perfect	Perfect
30	0.3P2112		Perfect	Perfect
31	0.3P2412		Perfect	Perfect
32	0.3P2512		Perfect	Perfect
33	0.3P3012		Perfect	Perfect
34	0.35P112		Perfect	Perfect
35	0.35P212		Perfect	Perfect

36	0.35P312	Perfect	Perfect
37	0.35P712	Perfect	Perfect
38	0.35P812	Perfect	Perfect
39	0.35P1512	Perfect	Perfect
40	0.35P1812	Perfect	Perfect
41	0.35P1912	Perfect	Perfect
42	0.35P2512	Perfect	Perfect
43	0.35P2812	Perfect	Perfect
44	0.35P2912	Perfect	Perfect
45	0.4P112	Perfect	Perfect
46	0.4P212	Perfect	Perfect
47	0.4P312	Perfect	Perfect
48	0.4P1512	Perfect	Perfect
49	0.4P1612	Perfect	Perfect
50	0.4P1712	Perfect	Perfect
51	0.4P1812	Perfect	Perfect
52	0.4P1912	Perfect	Perfect
53	0.4P2012	Perfect	Perfect
54	0.4P2512	Perfect	Perfect
55	0.4P2912	Perfect	Perfect
56	0.4P3012	Perfect	Perfect

Table A13

Closest Missing Data Theory Predictive Accuracy in Data Set 13 (DS13) Experiments for Integration Effort (IM) % - Reused Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (23)	MI-LR (33)
1	0.05P313			1
2	0.05P913		1	
3	0.1P213			1
4	0.1P913		1	
5	0.1P1813			1
6	0.15P313		1	
7	0.15P613			1
8	0.15P713		1	
9	0.15P2413		1	
10	0.15P3013		1	
11	0.2P213		1	
12	0.2P313			1
13	0.2P913		1	
14	0.2P1213		1	

15	0.2P1813	1	
16	0.2P2913	1	
17	0.25P613		1
18	0.25P813	1	
19	0.25P1113	1	
20	0.25P1213		1
21	0.25P1413		1
22	0.25P1713	1	
23	0.25P2513		1
24	0.25P2713	1	
25	0.3P113	1	
26	0.3P213	1	
27	0.3P1613		1
28	0.3P1813	1	
29	0.3P1913		1
30	0.3P2113		1
31	0.3P2413	1	
32	0.3P2513		1
33	0.3P3013		1
34	0.35P113		1
35	0.35P213		1
36	0.35P313		1
37	0.35P713	1	
38	0.35P813		1
39	0.35P1513		1
40	0.35P1813		1
41	0.35P1913		1
42	0.35P2513		1
43	0.35P2813		1
44	0.35P2913	1	
45	0.4P113		1
46	0.4P213		1
47	0.4P313		1
48	0.4P1513		1
49	0.4P1613		1
50	0.4P1713	1	
51	0.4P1813		1
52	0.4P1913		1
53	0.4P2013		1
54	0.4P2513		1
55	0.4P2913	1	
56	0.4P3013		1

Table A14

Closest Missing Data Theory Predictive Accuracy in Data Set 14 (DS14) Experiments for Final Software Requirements Analysis Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (30)	MI-LR (26)
1	0.05P314		1	
2	0.05P914			1
3	0.1P214			1
4	0.1P914			1
5	0.1P1814		1	
6	0.15P314			1
7	0.15P614			1
8	0.15P714			1
9	0.15P2414		1	
10	0.15P3014			1
11	0.2P214			1
12	0.2P314			1
13	0.2P914		1	
14	0.2P1214		1	
15	0.2P1814		1	
16	0.2P2914			1
17	0.25P614			1
18	0.25P814		1	
19	0.25P1114		1	
20	0.25P1214		1	
21	0.25P1414		1	
22	0.25P1714		1	
23	0.25P2514		1	
24	0.25P2714			1
25	0.3P114			1
26	0.3P214		1	
27	0.3P1614			1
28	0.3P1814			1
29	0.3P1914		1	
30	0.3P2114		1	
31	0.3P2414		1	
32	0.3P2514		1	
33	0.3P3014			1
34	0.35P114			1
35	0.35P214			1
36	0.35P314		1	
37	0.35P714		1	

38	0.35P814		1
39	0.35P1514	1	
40	0.35P1814	1	
41	0.35P1914	1	
42	0.35P2514	1	
43	0.35P2814	1	
44	0.35P2914		1
45	0.4P114		1
46	0.4P214		1
47	0.4P314		1
48	0.4P1514	1	
49	0.4P1614		1
50	0.4P1714	1	
51	0.4P1814	1	
52	0.4P1914	1	
53	0.4P2014	1	
54	0.4P2514	1	
55	0.4P2914		1
56	0.4P3014	1	1

Table A15

Closest Missing Data Theory Predictive Accuracy in Data Set 15 (DS15) Experiments for Final Software Architectural Design Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (23)	MI-LR (33)
1	0.05P315			1
2	0.05P915		1	
3	0.1P215			1
4	0.1P915		1	
5	0.1P1815		1	
6	0.15P315			1
7	0.15P615			1
8	0.15P715		1	
9	0.15P2415		1	
10	0.15P3015			1
11	0.2P215			1
12	0.2P315			1
13	0.2P915			1
14	0.2P1215			1
15	0.2P1815		1	
16	0.2P2915			1

17	0.25P615		1
18	0.25P815	1	
19	0.25P1115		1
20	0.25P1215	1	
21	0.25P1415	1	
22	0.25P1715	1	
23	0.25P2515	1	
24	0.25P2715		1
25	0.3P115		1
26	0.3P215		1
27	0.3P1615	1	
28	0.3P1815		1
29	0.3P1915	1	
30	0.3P2115	1	
31	0.3P2415	1	
32	0.3P2515	1	
33	0.3P3015		1
34	0.35P115		1
35	0.35P215		1
36	0.35P315		1
37	0.35P715		1
38	0.35P815		1
39	0.35P1515		1
40	0.35P1815		1
41	0.35P1915	1	
42	0.35P2515	1	
43	0.35P2815		1
44	0.35P2915		1
45	0.4P115		1
46	0.4P215		1
47	0.4P315		1
48	0.4P1515		1
49	0.4P1615	1	
50	0.4P1715	1	
51	0.4P1815		1
52	0.4P1915	1	
53	0.4P2015	1	
54	0.4P2515	1	
55	0.4P2915		1
56	0.4P3015		1

Table A16

Closest Missing Data Theory Predictive Accuracy in Data Set 16 (DS16) Experiments for Final Software Detailed Design Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (32)	MI-LR (24)
1	0.05P116			1
2	0.05P216		1	
3	0.05P316			1
4	0.05P416			1
5	0.05P516		1	
6	0.05P616		1	
7	0.05P716			1
8	0.05P816			1
9	0.05P916		1	
10	0.05P1016			1
11	0.05P1116			1
12	0.05P1216			1
13	0.05P1316		1	
14	0.05P1416		1	
15	0.05P1516		1	
16	0.05P1616			1
17	0.05P1716		1	
18	0.05P1816			1
19	0.05P1916		1	
20	0.05P2016			1
21	0.05P2116		1	
22	0.05P2216		1	
23	0.05P2316		1	
24	0.05P2416			1
25	0.05P2516		1	
26	0.05P2616		1	
27	0.05P2716		1	
28	0.05P2816		1	
29	0.05P2916		1	
30	0.05P3016		1	
31	0.05P116		1	
32	0.05P216		1	
33	0.05P316			1
34	0.05P416			1
35	0.05P516			1
36	0.05P616			1
37	0.05P716		1	

38	0.05P816	1	
39	0.05P916	1	
40	0.05P1016	1	
41	0.05P1116	1	
42	0.05P1216	1	
43	0.05P1316		1
44	0.05P1416		1
45	0.05P1516		1
46	0.05P1616		1
47	0.05P1716		1
48	0.05P1816	1	
49	0.05P1916	1	
50	0.05P2016		1
51	0.05P2116	1	
52	0.05P2216	1	
53	0.05P2316	1	
54	0.05P2416	1	
55	0.05P2516		1
56	0.05P2616		1

Table A17

Closest Missing Data Theory Predictive Accuracy in Data Set 17 (DS17) Experiments for Final Software Construction Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (32)	MI-LR (24)
1	0.05P317		1	
2	0.05P917		1	
3	0.1P217			1
4	0.1P917		1	
5	0.1P1817		1	
6	0.15P317			1
7	0.15P617			1
8	0.15P717			1
9	0.15P2417		1	
10	0.15P3017			1
11	0.2P217			1
12	0.2P317			1
13	0.2P917			1
14	0.2P1217		1	
15	0.2P1817			1
16	0.2P2917			1

17	0.25P617	1	
18	0.25P817		1
19	0.25P1117	1	
20	0.25P1217	1	
21	0.25P1417	1	
22	0.25P1717	1	
23	0.25P2517	1	
24	0.25P2717		1
25	0.3P117		1
26	0.3P217	1	
27	0.3P1617	1	
28	0.3P1817	1	
29	0.3P1917	1	
30	0.3P2117	1	
31	0.3P2417	1	
32	0.3P2517	1	
33	0.3P3017		1
34	0.35P117	1	
35	0.35P217	1	
36	0.35P317		1
37	0.35P717		1
38	0.35P817	1	
39	0.35P1517	1	
40	0.35P1817	1	
41	0.35P1917	1	
42	0.35P2517	1	
43	0.35P2817		1
44	0.35P2917		1
45	0.4P117		1
46	0.4P217		1
47	0.4P317		1
48	0.4P1517		1
49	0.4P1617	1	
50	0.4P1717	1	
51	0.4P1817	1	
52	0.4P1917	1	
53	0.4P2017	1	
54	0.4P2517	1	
55	0.4P2917		1
56	0.4P3017	1	1

Table A18

Closest Missing Data Theory Predictive Accuracy in Data Set 18 (DS18) Experiments for Final Software Integration Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (34)	MI-LR (22)
1	0.05P318			1
2	0.05P918			1
3	0.1P218			1
4	0.1P918		1	
5	0.1P1818		1	
6	0.15P318		1	
7	0.15P618		1	
8	0.15P718		1	
9	0.15P2418		1	
10	0.15P3018			1
11	0.2P218			1
12	0.2P318		1	
13	0.2P918		1	
14	0.2P1218		1	
15	0.2P1818		1	
16	0.2P2918			1
17	0.25P618			1
18	0.25P818		1	
19	0.25P1118		1	
20	0.25P1218		1	
21	0.25P1418		1	
22	0.25P1718		1	
23	0.25P2518			1
24	0.25P2718			1
25	0.3P118			1
26	0.3P218			1
27	0.3P1618		1	
28	0.3P1818		1	
29	0.3P1918		1	
30	0.3P2118		1	
31	0.3P2418		1	
32	0.3P2518		1	
33	0.3P3018			1
34	0.35P118			1
35	0.35P218		1	

36	0.35P318		1
37	0.35P718		1
38	0.35P818	1	
39	0.35P1518	1	
40	0.35P1818	1	
41	0.35P1918	1	
42	0.35P2518	1	
43	0.35P2818		1
44	0.35P2918		1
45	0.4P118		1
46	0.4P218		1
47	0.4P318		1
48	0.4P1518	1	
49	0.4P1618	1	
50	0.4P1718	1	
51	0.4P1818	1	
52	0.4P1918	1	
53	0.4P2018	1	
54	0.4P2518	1	
55	0.4P2918		1
56	0.4P3018		1

Table A19

Closest Missing Data Theory Predictive Accuracy in Data Set 19 (DS19) Experiments for Final Software Qualification Testing Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (29)	MI-LR (27)
1	0.05P319			1
2	0.05P919		1	
3	0.1P219		1	
4	0.1P919		1	
5	0.1P1819		1	
6	0.15P319			1
7	0.15P619			1
8	0.15P719			1
9	0.15P2419		1	
10	0.15P3019			1
11	0.2P219			1
12	0.2P319		1	
13	0.2P919			1

14	0.2P1219		1
15	0.2P1819	1	
16	0.2P2919		1
17	0.25P619	1	
18	0.25P819		1
19	0.25P1119		1
20	0.25P1219	1	
21	0.25P1419	1	
22	0.25P1719	1	
23	0.25P2519	1	
24	0.25P2719		1
25	0.3P119		1
26	0.3P219		1
27	0.3P1619	1	
28	0.3P1819	1	
29	0.3P1919	1	
30	0.3P2119	1	
31	0.3P2419	1	
32	0.3P2519	1	
33	0.3P3019		1
34	0.35P119		1
35	0.35P219	1	
36	0.35P319		1
37	0.35P719	1	
38	0.35P819		1
39	0.35P1519	1	
40	0.35P1819		1
41	0.35P1919	1	
42	0.35P2519	1	
43	0.35P2819		1
44	0.35P2919		1
45	0.4P119		1
46	0.4P219		1
47	0.4P319		1
48	0.4P1519	1	
49	0.4P1619	1	
50	0.4P1719		1
51	0.4P1819	1	
52	0.4P1919	1	
53	0.4P2019	1	
54	0.4P2519	1	
55	0.4P2919		1
56	0.4P3019		1

Table A20

Closest Missing Data Theory Predictive Accuracy in Data Set 20 (DS20) Experiments for Final Software Documentation Management Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (32)	MI-LR (24)
1	0.05P320			1
2	0.05P920			1
3	0.1P220			1
4	0.1P920		1	
5	0.1P1820		1	
6	0.15P320			1
7	0.15P620			1
8	0.15P720			1
9	0.15P2420		1	
10	0.15P3020			1
11	0.2P220			1
12	0.2P320			1
13	0.2P920		1	
14	0.2P1220		1	
15	0.2P1820		1	
16	0.2P2920			1
17	0.25P620		1	
18	0.25P820		1	
19	0.25P1120		1	
20	0.25P1220			1
21	0.25P1420		1	
22	0.25P1720		1	
23	0.25P2520		1	
24	0.25P2720			1
25	0.3P120		1	
26	0.3P220			1
27	0.3P1620		1	
28	0.3P1820		1	
29	0.3P1920		1	
30	0.3P2120			1
31	0.3P2420		1	
32	0.3P2520		1	
33	0.3P3020		1	
34	0.35P120		1	
35	0.35P220			1
36	0.35P320		1	

37	0.35P720		1
38	0.35P820		1
39	0.35P1520		1
40	0.35P1820		1
41	0.35P1920	1	
42	0.35P2520	1	
43	0.35P2820		1
44	0.35P2920		1
45	0.4P120	1	
46	0.4P220		1
47	0.4P320	1	
48	0.4P1520	1	
49	0.4P1620	1	
50	0.4P1720	1	
51	0.4P1820	1	
52	0.4P1920	1	
53	0.4P2020	1	
54	0.4P2520	1	
55	0.4P2920		1
56	0.4P3020		1

Table A21

Closest Missing Data Theory Predictive Accuracy in Data Set 21 (DS21) Experiments for Final Software Configuration Management Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (32)	MI-LR (24)
1	0.05P321			1
2	0.05P921		1	
3	0.1P221			1
4	0.1P921		1	
5	0.1P1821			1
6	0.15P321			1
7	0.15P621			1
8	0.15P721			1
9	0.15P2421		1	
10	0.15P3021			1
11	0.2P221		1	
12	0.2P321			1
13	0.2P921		1	
14	0.2P1221			1
15	0.2P1821			1

16	0.2P2921		1
17	0.25P621		1
18	0.25P821		1
19	0.25P1121	1	
20	0.25P1221	1	
21	0.25P1421		1
22	0.25P1721	1	
23	0.25P2521	1	
24	0.25P2721		1
25	0.3P121	1	
26	0.3P221	1	
27	0.3P1621	1	
28	0.3P1821	1	
29	0.3P1921	1	
30	0.3P2121	1	
31	0.3P2421	1	
32	0.3P2521	1	
33	0.3P3021	1	
34	0.35P121	1	
35	0.35P221	1	
36	0.35P321		1
37	0.35P721		1
38	0.35P821	1	
39	0.35P1521	1	
40	0.35P1821		1
41	0.35P1921	1	
42	0.35P2521	1	
43	0.35P2821		1
44	0.35P2921		1
45	0.4P121	1	
46	0.4P221	1	
47	0.4P321		1
48	0.4P1521	1	
49	0.4P1621	1	
50	0.4P1721	1	
51	0.4P1821	1	
52	0.4P1921	1	
53	0.4P2021		1
54	0.4P2521	1	
55	0.4P2921		1
56	0.4P3021		1

Table A22

Closest Missing Data Theory Predictive Accuracy in Data Set 22 (DS22) Experiments for Final Software Quality Assurance Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (31)	MI-LR (25)
1	0.05P322			1
2	0.05P922			1
3	0.1P222			1
4	0.1P922		1	
5	0.1P1822		1	
6	0.15P322			1
7	0.15P622		1	
8	0.15P722		1	
9	0.15P2422		1	
10	0.15P3022			1
11	0.2P222		1	
12	0.2P322		1	
13	0.2P922		1	
14	0.2P1222		1	
15	0.2P1822		1	
16	0.2P2922			1
17	0.25P622			1
18	0.25P822			1
19	0.25P1122		1	
20	0.25P1222			1
21	0.25P1422		1	
22	0.25P1722		1	
23	0.25P2522		1	
24	0.25P2722			1
25	0.3P122			1
26	0.3P222			1
27	0.3P1622		1	
28	0.3P1822		1	
29	0.3P1922		1	
30	0.3P2122		1	
31	0.3P2422		1	
32	0.3P2522		1	
33	0.3P3022			1
34	0.35P122			1
35	0.35P222			1
36	0.35P322			1
37	0.35P722			1

38	0.35P822		1
39	0.35P1522	1	
40	0.35P1822	1	
41	0.35P1922	1	
42	0.35P2522	1	
43	0.35P2822	1	
44	0.35P2922		1
45	0.4P122		1
46	0.4P222		1
47	0.4P322		1
48	0.4P1522	1	
49	0.4P1622	1	
50	0.4P1722	1	
51	0.4P1822	1	
52	0.4P1922	1	
53	0.4P2022	1	
54	0.4P2522		1
55	0.4P2922		1
56	0.4P3022		1

Table A23

Closest Missing Data Theory Predictive Accuracy in Data Set 23 (DS23) Experiments for Final Software Verification Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (23)	MI-LR (33)
1	0.05P323			1
2	0.05P923			1
3	0.1P223			1
4	0.1P923			1
5	0.1P1823		1	
6	0.15P323			1
7	0.15P623			1
8	0.15P723			1
9	0.15P2423		1	
10	0.15P3023			1
11	0.2P223			1
12	0.2P323			1
13	0.2P923			1
14	0.2P1223		1	
15	0.2P1823			1
16	0.2P2923			1

17	0.25P623		1
18	0.25P823		1
19	0.25P1123	1	
20	0.25P1223		1
21	0.25P1423	1	
22	0.25P1723	1	
23	0.25P2523	1	
24	0.25P2723		1
25	0.3P123		1
26	0.3P223		1
27	0.3P1623		1
28	0.3P1823	1	
29	0.3P1923	1	
30	0.3P2123	1	
31	0.3P2423	1	
32	0.3P2523	1	
33	0.3P3023		1
34	0.35P123		1
35	0.35P223		1
36	0.35P323		1
37	0.35P723		1
38	0.35P823		1
39	0.35P1523	1	
40	0.35P1823	1	
41	0.35P1923	1	
42	0.35P2523	1	
43	0.35P2823		1
44	0.35P2923		1
45	0.4P123		1
46	0.4P223		1
47	0.4P323		1
48	0.4P1523	1	
49	0.4P1623	1	
50	0.4P1723	1	
51	0.4P1823	1	
52	0.4P1923	1	
53	0.4P2023	1	
54	0.4P2523	1	
55	0.4P2923		1
56	0.4P3023		1

Table A24

Closest Missing Data Theory Predictive Accuracy in Data Set 24 (DS24) Experiments for Final Software Validation Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (35)	MI-LR (21)
1	0.05P324		1	
2	0.05P924			1
3	0.1P224		1	
4	0.1P924		1	
5	0.1P1824		1	
6	0.15P324			1
7	0.15P624			1
8	0.15P724			1
9	0.15P2424		1	
10	0.15P3024			1
11	0.2P224		1	
12	0.2P324		1	
13	0.2P924		1	
14	0.2P1224		1	
15	0.2P1824		1	
16	0.2P2924			1
17	0.25P624			1
18	0.25P824			1
19	0.25P1124		1	
20	0.25P1224		1	
21	0.25P1424		1	
22	0.25P1724		1	
23	0.25P2524		1	
24	0.25P2724		1	
25	0.3P124		1	
26	0.3P224		1	
27	0.3P1624		1	
28	0.3P1824		1	
29	0.3P1924			1
30	0.3P2124		1	
31	0.3P2424		1	
32	0.3P2524		1	
33	0.3P3024			1
34	0.35P124			1
35	0.35P224		1	
36	0.35P324			1
37	0.35P724		1	

38	0.35P824	1	
39	0.35P1524	1	
40	0.35P1824	1	
41	0.35P1924	1	
42	0.35P2524	1	
43	0.35P2824		1
44	0.35P2924		1
45	0.4P124		1
46	0.4P224		1
47	0.4P324		1
48	0.4P1524	1	
49	0.4P1624	1	
50	0.4P1724		1
51	0.4P1824	1	
52	0.4P1924		1
53	0.4P2024	1	
54	0.4P2524	1	
55	0.4P2924		1
56	0.4P3024		1

Table A25

Closest Missing Data Theory Predictive Accuracy in Data Set 25 (DS25) Experiments for Final Software Review Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (34)	MI-LR (22)
1	0.05P325		1	
2	0.05P925		1	
3	0.1P225		1	
4	0.1P925		1	
5	0.1P1825		1	
6	0.15P325		1	
7	0.15P625		1	
8	0.15P725		1	
9	0.15P2425			1
10	0.15P3025			1
11	0.2P225		1	
12	0.2P325			1
13	0.2P925		1	
14	0.2P1225			1
15	0.2P1825		1	

16	0.2P2925		1
17	0.25P625	1	
18	0.25P825	1	
19	0.25P1125	1	
20	0.25P1225	1	
21	0.25P1425	1	
22	0.25P1725	1	
23	0.25P2525	1	
24	0.25P2725	1	
25	0.3P125		1
26	0.3P225	1	
27	0.3P1625	1	
28	0.3P1825	1	
29	0.3P1925	1	
30	0.3P2125	1	
31	0.3P2425	1	
32	0.3P2525	1	
33	0.3P3025		1
34	0.35P125		1
35	0.35P225		1
36	0.35P325		1
37	0.35P725		1
38	0.35P825		1
39	0.35P1525	1	
40	0.35P1825	1	
41	0.35P1925	1	
42	0.35P2525	1	
43	0.35P2825		1
44	0.35P2925		1
45	0.4P125		1
46	0.4P225		1
47	0.4P325		1
48	0.4P1525		1
49	0.4P1625	1	
50	0.4P1725		1
51	0.4P1825	1	
52	0.4P1925	1	
53	0.4P2025	1	
54	0.4P2525		1
55	0.4P2925		1
56	0.4P3025		1

Table A26

Closest Missing Data Theory Predictive Accuracy in Data Set 26 (DS26) Experiments for Final Software Audit Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (28)	MI-LR (28)
1	0.05P326		1	
2	0.05P926			1
3	0.1P226			1
4	0.1P926			1
5	0.1P1826			1
6	0.15P326		1	
7	0.15P626		1	
8	0.15P726			1
9	0.15P2426		1	
10	0.15P3026			1
11	0.2P226			1
12	0.2P326			1
13	0.2P926		1	
14	0.2P1226		1	
15	0.2P1826			1
16	0.2P2926			1
17	0.25P626		1	
18	0.25P826		1	
19	0.25P1126		1	
20	0.25P1226		1	
21	0.25P1426			1
22	0.25P1726			1
23	0.25P2526		1	
24	0.25P2726		1	
25	0.3P126		1	
26	0.3P226			1
27	0.3P1626			1
28	0.3P1826			1
29	0.3P1926		1	
30	0.3P2126		1	
31	0.3P2426		1	
32	0.3P2526		1	
33	0.3P3026			1
34	0.35P126		1	
35	0.35P226			1
36	0.35P326			1
37	0.35P726			1

38	0.35P826	1	
39	0.35P1526	1	
40	0.35P1826		1
41	0.35P1926		1
42	0.35P2526	1	
43	0.35P2826		1
44	0.35P2926		1
45	0.4P126	1	
46	0.4P226		1
47	0.4P326		1
48	0.4P1526	1	
49	0.4P1626	1	
50	0.4P1726	1	
51	0.4P1826		1
52	0.4P1926	1	
53	0.4P2026	1	
54	0.4P2526	1	
55	0.4P2926		1
56	0.4P3026	1	1

Table A27

Closest Missing Data Theory Predictive Accuracy in Data Set 27 (DS27) Experiments for Final Software Problem Resolution Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (25)	MI-LR (31)
1	0.05P327			1
2	0.05P927			1
3	0.1P227			1
4	0.1P927		1	
5	0.1P1827			1
6	0.15P327			1
7	0.15P627			1
8	0.15P727			1
9	0.15P2427		1	
10	0.15P3027			1
11	0.2P227			1
12	0.2P327			1
13	0.2P927		1	
14	0.2P1227		1	
15	0.2P1827			1

16	0.2P2927		1
17	0.25P627		1
18	0.25P827		1
19	0.25P1127		1
20	0.25P1227	1	
21	0.25P1427	1	
22	0.25P1727	1	
23	0.25P2527	1	
24	0.25P2727	1	
25	0.3P127		1
26	0.3P227		1
27	0.3P1627	1	
28	0.3P1827		1
29	0.3P1927	1	
30	0.3P2127	1	
31	0.3P2427	1	
32	0.3P2527	1	
33	0.3P3027		1
34	0.35P127		1
35	0.35P227		1
36	0.35P327		1
37	0.35P727	1	
38	0.35P827	1	
39	0.35P1527	1	
40	0.35P1827		1
41	0.35P1927	1	
42	0.35P2527	1	
43	0.35P2827	1	
44	0.35P2927		1
45	0.4P127		1
46	0.4P227		1
47	0.4P327		1
48	0.4P1527		1
49	0.4P1627	1	
50	0.4P1727	1	
51	0.4P1827		1
52	0.4P1927	1	
53	0.4P2027	1	
54	0.4P2527	1	
55	0.4P2927		1
56	0.4P3027		1

Table A28

Closest Missing Data Theory Predictive Accuracy in Data Set 28 (DS28) Experiments for Final Cybersecurity Effort Hours Types of Data

Id	Experiment Run Case No.	LD (0)	SI-Mean (54)	MI-LR (2)
1	0.05P328		1	
2	0.05P928		1	
3	0.1P228		1	
4	0.1P928		1	
5	0.1P1828		1	
6	0.15P328		1	
7	0.15P628			1
8	0.15P728		1	
9	0.15P2428		1	
10	0.15P3028		1	
11	0.2P228		1	
12	0.2P328		1	
13	0.2P928		1	
14	0.2P1228		1	
15	0.2P1828		1	
16	0.2P2928		1	
17	0.25P628			1
18	0.25P828		1	
19	0.25P1128		1	
20	0.25P1228		1	
21	0.25P1428		1	
22	0.25P1728		1	
23	0.25P2528		1	
24	0.25P2728		1	
25	0.3P128		1	
26	0.3P228		1	
27	0.3P1628		1	
28	0.3P1828		1	
29	0.3P1928		1	
30	0.3P2128		1	
31	0.3P2428		1	
32	0.3P2528		1	
33	0.3P3028		1	
34	0.35P128		1	
35	0.35P228		1	
36	0.35P328		1	
37	0.35P728		1	

38	0.35P828	1
39	0.35P1528	1
40	0.35P1828	1
41	0.35P1928	1
42	0.35P2528	1
43	0.35P2828	1
44	0.35P2928	1
45	0.4P128	1
46	0.4P228	1
47	0.4P328	1
48	0.4P1528	1
49	0.4P1628	1
50	0.4P1728	1
51	0.4P1828	1
52	0.4P1928	1
53	0.4P2028	1
54	0.4P2528	1
55	0.4P2928	1
56	0.4P3028	1

Appendix B: Two-Way Repeated Measures ANOVA in SPSS Selection

Figure B1

Select Analyze, General Linear Model, and Repeated Measures Screen

Repeated Measures Define Factor(s) X

Within-Subject Factor Name: factor1

Number of Levels: 1

Add Change Remove

Error(2)

Measure Name:

Add Change Remove

Define Reset Cancel Help

Figure B2

Select Within-Subjects Variables and Between-Subjects Factors

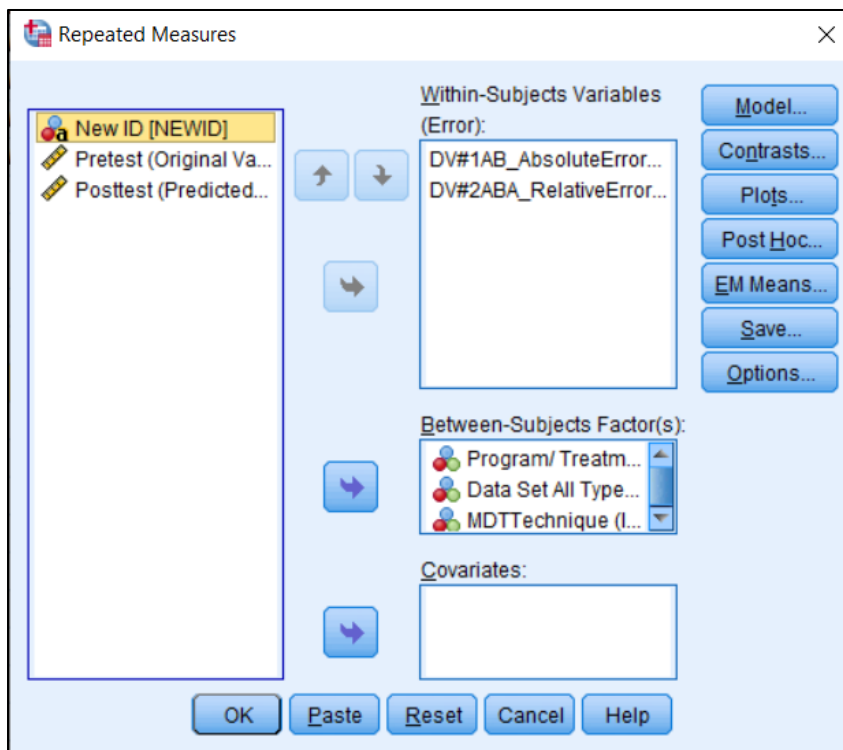
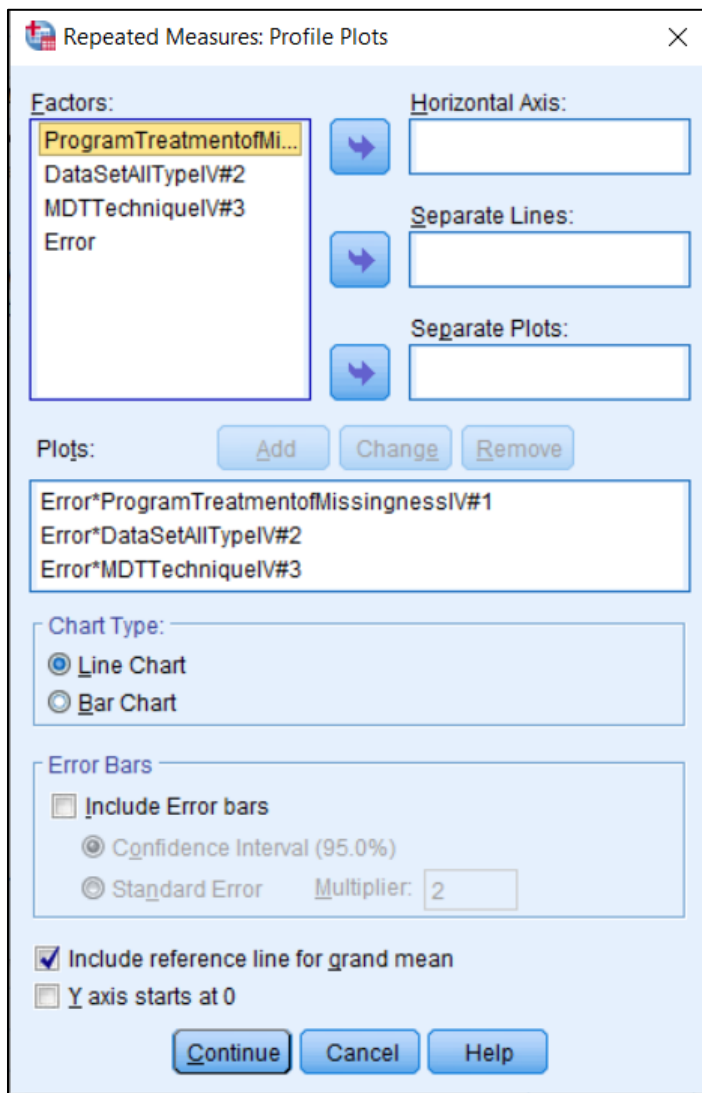


Figure B3

Define Profile Plots to Determine if the Means are Equal on each Missing Data Theory Technique



The image shows the 'Repeated Measures: Profile Plots' dialog box in SPSS. The 'Factors' list on the left contains 'ProgramTreatmentofMi...', 'DataSetAllTypeIV#2', 'MDTTechniqueIV#3', and 'Error'. The 'Horizontal Axis' field is empty. The 'Separate Lines' and 'Separate Plots' fields are also empty. The 'Plots' section contains 'Error*ProgramTreatmentofMissingnessIV#1', 'Error*DataSetAllTypeIV#2', and 'Error*MDTTechniqueIV#3'. The 'Chart Type' section has 'Line Chart' selected. The 'Error Bars' section has 'Include Error bars' checked, with 'Confidence Interval (95.0%)' selected and a 'Multiplier' of 2. The 'Include reference line for grand mean' checkbox is checked, and 'Y axis starts at 0' is unchecked. Buttons for 'Continue', 'Cancel', and 'Help' are at the bottom.

Repeated Measures: Profile Plots

Factors:

- ProgramTreatmentofMi...
- DataSetAllTypeIV#2
- MDTTechniqueIV#3
- Error

Horizontal Axis:

Separate Lines:

Separate Plots:

Plots: Add Change Remove

- Error*ProgramTreatmentofMissingnessIV#1
- Error*DataSetAllTypeIV#2
- Error*MDTTechniqueIV#3

Chart Type:

- Line Chart
- Bar Chart

Error Bars

- Include Error bars
 - Confidence Interval (95.0%)
 - Standard Error Multiplier: 2

Include reference line for grand mean

Y axis starts at 0

Continue Cancel Help

Figure B4

Define Post Hoc Tests for the Independent Variables

Repeated Measures: Post Hoc Multiple Comparisons for Observed Means

Factor(s):
ProgramTreatmentofMissingnes...
DataSetAllTypeIV#2
MDTTechniqueIV#3

Post Hoc Tests for:
ProgramTreatmentofMissingnes...
DataSetAllTypeIV#2
MDTTechniqueIV#3

Equal Variances Assumed

LSD S-N-K Waller-Duncan
 Bonferroni Tukey Type I/Type II Error Ratio: 100
 Sidak Tukey's-b Dunnnett
 Scheffe Duncan Control Category: Last
 R-E-G-W-F Hochberg's GT2 Test
 R-E-G-W-Q Gabriel 2-sided < Control > Control

Equal Variances Not Assumed

Tamhane's T2 Dunnnett's T3 Games-Howell Dunnnett's C

Continue Cancel Help

Figure B5

Define Estimated Marginal Means

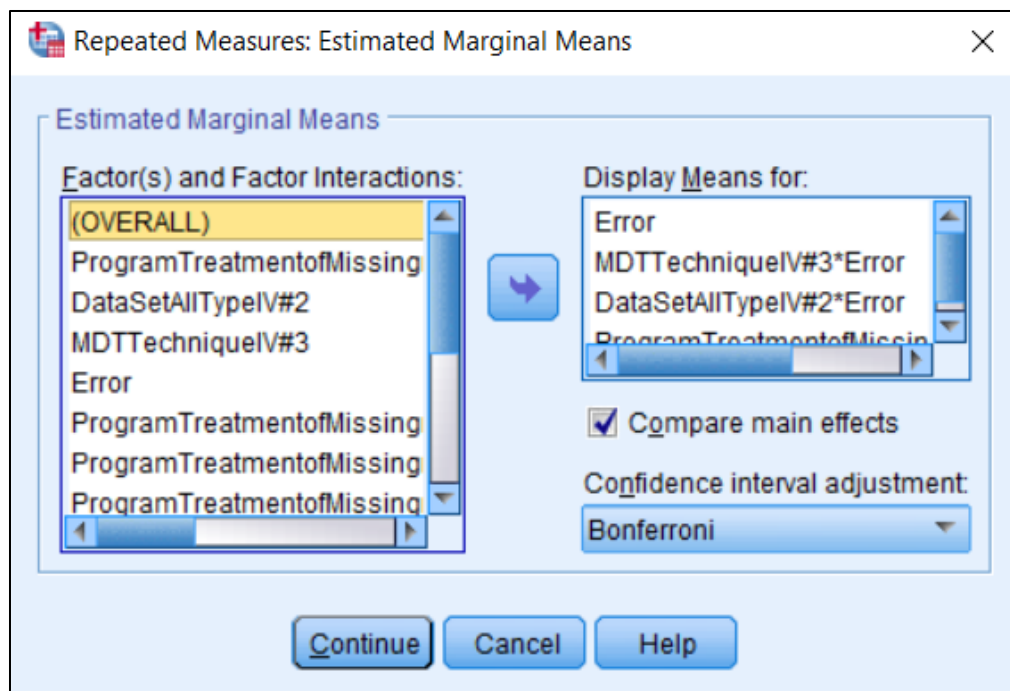


Figure B6*Define Options to Analyze*

Repeated Measures: Options

Display

- Descriptive statistics
- Estimates of effect size
- Observed power
- Parameter estimates
- SSCP matrices
- Residual SSCP matrix
- Transformation matrix
- Homogeneity tests
- Spread vs. level plot
- Residual plot
- Lack of fit
- General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Continue Cancel Help

Appendix C: Select SPSS Outputs from Two-Way Repeated Measures ANOVA

Figure C1*Within-Subjects Factors Coded in SPSS*

Within-Subjects Factors
Measure: MEASURE_1

Pretest_Posttest	Dependent Variable
1	PretestOriginalValue
2	PosttestPredictedValue

Figure C2*Between-Subjects Factors Coded in SPSS*

Between-Subjects Factors

		N
Program/ Treatment % of Missingness (IV 1)	5	112
	10	168
	15	280
	20	336
	25	448
	30	504
	35	616
	40	672
	Data Set All Type (IV 2)	1
2		112
3		112
4		112
5		112
6		112
7		112
8		112

	9	112
	10	112
	11	112
	12	112
	13	112
	14	112
	15	112
	16	112
	17	112
	18	112
	19	112
	20	112
	21	112
	22	112
	23	112
	24	112
	25	112
	26	112
	27	112
	28	112
MDTTechnique (IV 3)	2	1568
	3	1568

Figure C3*Box's Test of Equality of Covariance Matrices***Box's Test of
Equality of
Covariance
Matrices^a**

Box's M	21757.008
F	34.467
df1	528
df2	30316.769
Sig.	.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design:

Intercept +
 ProgramTreat
 mentofMissin
 gnessIV#1 +
 DataSetAllTyp
 eIV#2 +
 MDTTechniqu
 eIV#3 +
 ProgramTreat
 mentofMissin
 gnessIV#1 *
 DataSetAllTyp
 eIV#2 +
 ProgramTreat
 mentofMissin
 gnessIV#1 *
 MDTTechniqu
 eIV#3 +
 DataSetAllTyp
 eIV#2 *
 MDTTechniqu
 eIV#3 +
 ProgramTreat
 mentofMissin
 gnessIV#1 *
 DataSetAllTyp
 eIV#2 *
 MDTTechniqu
 eIV#3
 Within
 Subjects
 Design:
 Pretest_Postt
 est



Figure C4

Mauchly's Test of Sphericity

Mauchly's Test of Sphericity^a

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^b		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Pretest_Posttest	1.000	.000	0	.	1.000	1.000	1.000

→ Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept + ProgramTreatmentofMissingnessIV#1 + DataSetAllTypeIV#2 + MDTTechniqueIV#3 + ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2 + ProgramTreatmentofMissingnessIV#1 * MDTTechniqueIV#3 + DataSetAllTypeIV#2 * MDTTechniqueIV#3 + ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2 * MDTTechniqueIV#3
 Within Subjects Design: Pretest_Posttest

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Figure C5

Levene's Test of Equality of Error Variances

Levene's Test of Equality of Error Variances^a

		Levene Statistic	df1	df2	Sig.
Pretest (Original Value)	Based on Mean	5.138	447	2688	.000
	Based on Median	1.647	447	2688	.000
	Based on Median and with adjusted df	1.647	447	83.596	.003
	Based on trimmed mean	3.678	447	2688	.000
Posttest (Predicted Value)	Based on Mean	12.695	447	2688	.000
	Based on Median	6.472	447	2688	.000
	Based on Median and with adjusted df	6.472	447	66.584	.000
	Based on trimmed mean	11.893	447	2688	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + ProgramTreatmentofMissingnessIV#1 + DataSetAllTypeIV#2 + MDTTechniqueIV#3 + ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2 + ProgramTreatmentofMissingnessIV#1 * MDTTechniqueIV#3 + DataSetAllTypeIV#2 * MDTTechniqueIV#3 + ProgramTreatmentofMissingnessIV#1 * DataSetAllTypeIV#2 * MDTTechniqueIV#3
 Within Subjects Design: Pretest_Posttest