DISSERTATION

ANALYTICAL INJUSTICE LEAGUE: UNDERSTANDING STATISTICAL MANIPULATION OF STUDENT RETENTION DATA USING MODIFICATION METHODS FOR MISSING VALUES

Submitted by

Sarah E. Long

School of Education

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2021

Doctoral Committee:

Advisor: Gene Gloeckner

Sharon Anderson
James Folkestad
Aaron Eakman

ABSTRACT

ANALYTICAL INJUSTICE LEAGUE: UNDERSTANDING STATISTICAL MANIPULATION OF STUDENT RETENTION DATA USING MODIFICATION METHODS FOR MISSING VALUES

Missing values that fail to be appropriately accounted for may lead to reduced statistical power, biased estimators, reduced representativeness of the sample, and incorrect interpretations and conclusions (Gorelick, 2006). The current study provided an ontological perspective of data manipulation by explaining how statistical results can fundamentally change depending on specific data modification methods. This has consequential implications, specifically in higher education, that depend on quantifiable methodologies to substantiate practices through evidence-based policy making (Gillborn et al., 2018; Sindhi et al., 2019). The results of the current study exposed how examining patterns of data missingness can have critical implications on student retention initiatives including intervention programs, identification of high-risk students, and funding opportunities for support programs. It is imperative for both data scientists and data stakeholders to be critically aware of what data they collect, report, and utilize from the variable selection to statistical methodologies.

ACKNOWLEDGEMENTS

I want to first thank my wonderful partner, best friend in life, and soon to be husband, Stephen Cucchiara, for his unconditional love and support through this process. As I begin to write my wedding vows and conclude my doctoral dissertation, I can't help but think how truly grateful I am to have you in my life and how this journey began and ended with you by my side. It is not without saying that this one is for you…for us. Thank you. I want to thank my father, Bill Long, who always told me to go for it and who encouraged me to take a risk and move to colorful Colorado and begin this wild adventure. Although we are thousands of miles apart, you have been with me this entire time, pushing me to move forward, and telling me it will be worth it. You are my hero and role model. I also want to thank my beautiful mother, Becky Long, and sister, Ann Marie Angus. Thank you for reminding me that I come from a family of fearless women whose strength perseveres over all obstacles and challenges that come her way.

Big thank you to my advisor and overall wonderful human being, Gene Gloeckner, who told me all the things I needed to hear (even when I didn't want to). You made me feel valued as a scholar. I cannot thank you enough for that. To Sharon Anderson, who let me cry on her office couch many times but never once doubted my abilities. To James Folkestad, who took a chance on me and accepted me into the program and stuck with me ever since. To my outside committee members: The past: Victoria Buchan, and Jerry Vaske, and the present: Aaron Eakman. Thank you for being with me along this journey. Lastly, to the wonderful people I have met along the way: Megan Huwa (AMAZING editor and chief formatting officer- literally no words for your magic), Gene's advising seminar, for being cognitive at 7AM through feedback and positive energy. Dr. Elizabeth Jach and Dr. Susan Ernst, my PhD role models. I am so grateful for you.

DEDICATION

For Stephen and kitties

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: OVERVIEW OF THE STUDY

In an era of information overload, data are continuously being collected to track, monitor, and influence individuals' daily lives, from their political preferences to their entertainment accounts (Golbeck, 2016). The data gathered are used to explore trends, develop new products, and optimize user experience through web-based tracking and targeted advertisements. Some data are even used to score consumers based on their behavioral patterns (i.e., FICO Safe Driving Score; Cadwalladr & Graham-Harrison, 2018). Search engines, social media platforms, and other websites collect raw data (e.g., shares, likes, clicks, web searches, location) and generate computer based online profiles about each of their users. This digital persona can then predict the demographics, interests, and preferences of the consumer. For instance, Hill (2012) reported on a story where major store retailer, Target, used a prediction model to infer one of their shoppers was pregnant based on their online buying and viewing patterns which led to the shopper receiving mail-based coupons for those items to their home. It was later discovered that this person was a high-school student whose parents were not aware of their pregnancy until they began receiving coupons for baby clothes and cribs. Target did not hack into this individual's personally identifiable information (PPI), but instead based their conclusions on the 25 items that are correlated with pregnant shoppers (Hill, 2012). This example shows how shopping habits, as well as other information, are tracked to create new sets of data points that bypass direct information collection methods and are then used to predict behavior and build an entire profile without the individual's knowledge or consent (Tuttle, 2018).

In higher education, institutions are also collecting, analyzing, and profiling targeted user information in the form of student data. Though this information is not related to shopping

habits, data gathered from enrollment demographics, learning management systems, co-curricular programming, facility usage, and other aspects of student engagement are being used to provide evidence regarding retention, persistence, and graduation which serve as important indicators of overall student satisfaction as well as key performance measurements of the university (Daniel, 2015; Papamitsiou & Economides, 2014; Picciano, 2012; Williamson, 2017). These data are compiled for reporting to state and federal stakeholders as well as provide information to university administration to make strategic data-driven decisions (Rajuladevi, 2018; Williamson, 2017). These compilations of aggregated data are used to develop policies, procedures, and initiatives from macro levels (accreditation and funding) to micro levels (program development for targeted student populations; Hagood, 2019).

Statement of the Problem

In higher education, student success strategies, such as proactive interventions for at-risk populations, are often developed using an accumulation of data gathered from various sources. Information is then imputed into datasets that are used to build categories of learning behavior and predict a student's risk level of failing or leaving the university. Utilizing disaggregated data is usually not done due to student privacy concerns as well as the sheer volume of data needing to be analyzed (Fike & Fike, 2008). Therefore, inferences are often made using a plethora of information compiled from sources which often contain missing, inaccurate, or incomplete records (Cox et al., 2014).

A high number of values that are missing when data are collected and reported are a common, yet severe, problem that is often overlooked in quantitative educational research (Little et al., 2016). Missing values that fail to be appropriately accounted for may lead to reduced

statistical power, biased estimators, reduced representativeness of the sample, and incorrect interpretations and conclusions (Gorelick, 2006). It is often caused by students being reluctant to answer specific questions (e.g., income, ethnicity), data or coding entry errors, or information not being collected or reported (e.g., data privacy, security). The gravity of missing data comes from how much of the dataset is missing, source of missingness, and how unaccounted values are inputted into the observed dataset (Cox et al., 2014; Vaske, 2008). These factors "determine the magnitude of bias in estimates" (Vaske, 2008, p. 533); meaning, if there are large portions of missing data, a clear pattern for missingness that is not random (to be discussed in subsequent chapters), or methods of missing data imputation that discard or alter target populations, there are causes for concerns regarding the integrity of the dataset and therefore the implications of the results (Little et al., 2016). In higher education, this may mean unintentional discrimination against specific populations based on skewed results or missing vital student information that could have prevented attrition (Enders & Baraldi, 2018).

For example, data collected from the Integrated Postsecondary Education Data System (IPEDS), a federal agency designed to collect and report on post-secondary educational data, are "calculated based on data from first-time, full-time freshman students who graduate within six years of their initial enrollment date" (Yu et al., 2010, p. 308). Students who do not fit this criterion (e.g., attending part-time, non-traditional) are not accounted for in the data, and therefore, their characteristics are not evaluated for proactive interventions and policies (Fike & Fike, 2008). Since 2008, IPEDS has expanded their inclusion of retention criteria to encompass students attending part-time, as well as included demographic information. However, data collected in these areas are based on self-reports where students are given the option whether to disclose this information or not (Fuller, 2011). Information collected from IPEDS and other

governmental agencies are essential due to their role in the establishment of national benchmarks and performance-based assessments. Therefore, data accuracy is imperative not only for funding opportunities but also in the identification of best practices for student success (Zlatkin-Troitschanskaia et al., 2018).

In sum, although data are constantly being collected and used to make assumptions about its users, there is a significant amount of information that is not being collected or is missing within the dataset being analyzed. The way a researcher or data scientist decides the method for handling missing data over another is usually based on preference, research design, or other inclinations that can alter which variables are used or omitted in the final analysis (Leahey, 2008). This form of data modification has consequential implications on output interpretation which directly impacts various fields of study, specifically higher education, that depend on quantifiable methodologies to substantiate practices through evidence-based policy making (Gillborn et al., 2018; Sindhi et al., 2019). Therefore, the need to critically interrogate potential data manipulation needs to be investigated through the examination of statistical analysis, the reasons for missing data, and missing data patterns.

Research Significance

This study specifically examined the factors that impact student retention in higher education, which has been a longstanding benchmark of success for universities. Students withdrawing from an institution prior to program completion not only impacts the student but their family, the educational system, and the greater community they are a part of (Crosling, 2017). Further, as student profiles become increasingly more diverse, the need to "provide educational processes and programs that are inclusive for all students" (Crosling, 2017, p. 1) has become priority to most universities. This includes a shared responsibility between students and

the campus in ensuring students have the tools to be academically successful and persist to graduation. Through the collection and analysis of current and perspective student data, universities are able to gather data rich information to mitigate student attrition and provide proactive evidence-based interventions through early-alert systems and monitoring student engagement in real time (Picciano, 2012; Rajuladevi, 2018).

However, the sheer volume of data collected about students that can be inputted into a prediction model is far too large to be analyzed within one dataset. Therefore, data scientists often rely on previous research regarding factors that impact student retention (Aljohani, 2016; Rizkallah & Seitz, 2017). These factors, which are yielded as having the highest prediction of retention according to the Integrated Postsecondary Education Data System (2020), include student demographics, employment status, credit load, campus engagement, high school GPA, standardized test scores (SAT/ACT), current college GPA, major, and financial aid. Understanding what variables are used within an observed dataset provides stakeholders, such as researchers, data scientists, and administration, a clearer picture of the scope of the study, and the ability to express inequities due to the lack of specific data collected or omitted.

The current study used data from the University of Colorado Colorado Springs (UCCS) where there was a significant amount of data missing from student retention categories. For example, in the Demographic Information Fall 2019 Cohort from First-Year Cohort Retention Report (University of Colorado Colorado Springs [UCCS], 2020b), a total of 1,787 students reported their gender, race, or ethnicity but only 1,391 reported their estimated income. Additionally, 1,788 students reported their high school GPA; whereas over 1,250 did not report their ACT scores and over 200 did not report SAT scores. Although this is a snapshot of the

factors used to determine retention; the varying samples are cause for concern because of potential discrepancies in the data.

Researchers (Dalton & Thatcher, 2014; Iliadis & Russo, 2016) have grappled with problems such as using variables that have significant amounts of missing data but historically yield high prediction of the measurable outcomes: in this case, ACT scores and income. Questions about excluding missing variables or inputting the data through various methods have often been discussed; however, if variables regarding specificstudent populations are missing or omitted from the dataset, there is potential bias in the statistical output and the initiatives that may be developed using said data.

Close examination of methodological data mining techniques is not frequently explored in the literature (Aliyeva et al., 2018; Elish & Boyd, 2016; Knox, 2017). Often, researchers use their own judgement and ideations to either eliminate, substitute, or predict unaccounted data, which can change the outcome of the results (Staw, 1981). The current study provided insight into the methodological techniques of educational research by unpacking the implications of missing student retention data. This research filled the gap between the "acquisition of data and its use to advance discovery and innovation" (McNeely & Hahm, 2014, p. 304) by interrogating the opaqueness of the black box data methods through the understanding of the practical applications and deeper contextual insights of statistical methodologies (Boyd & Crawford, 2012; Yousif, 2015).

Operational Definitions

- *Retention* is categorized as students who are continually enrolled within the same institution from fall in their first year to the fall in their second year (National Student Clearinghouse Research Center, 2018).

- *Persistence* is defined as students who continue enrollment at any higher education institution for their second year (National Student Clearinghouse Research Center, 2018).
- *Data mining* is the analysis of patterns used to create the predictive model (Nyce, 2007).
- *Data missing completely at random (MCAR*) refers to missing data unrelated to the person/variable being studied. For example, a questionnaire was not filled out because it was lost in the mail, or a blood sample is missing because it was damaged in the lab (Peugh & Enders, 2004). The missingness of the item cannot be attributed to the individual but to outside circumstances not related to what is being studied (Meeyai, 2016).
- *Data missing at random (MAR*) signifies that the missingness is not connected to the person/variable being studied but is related to something in the dataset (Graham et al., 1996). For example, Bland (2015) explained:

  > If a child does not attend an educational assessment because the child is (genuinely) ill, this might be predictable from other data we have about the child's health, but it would not be related to what we would have measured had the child not been ill. (p. 306)

- *Data missing not at random (MNAR)* are missing variables directly related to what is being measured. For example, students who perform poorly on reading achievement exams are more likely to skip class on the day reading is being tested. Another example would be undocumented individuals not responding to census question about citizenship due to fear of prosecution.

Study Overview

Previous literature has addressed the factors that impact student retention as well as methods to mitigate missing values within a dataset; however, there was little known as to how missing data techniques impact the implications of student retention analysis. A conceptual map showing the gap in knowledge can be seen in Figure 1.1.

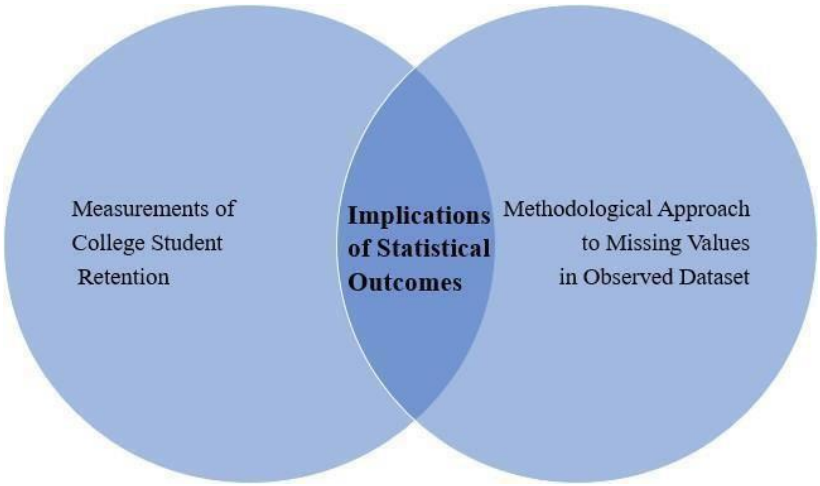


Figure 1.1

Conceptual Map of Study Overview

Therefore, the following overarching research question directed this manuscript: Howdoes using different methodological techniques for missing values in each dataset impact the statistical outcome? To address this research, I explored the following questions:

1. How differently does the utilization of missing data techniques predict student retention using a combination of student demographics, employment status, credit load, campus engagement, high school GPA, standardized test scores (SAT/ACT), current college GPA,major, and financial aid status?
    a. Which of the above variables are the best predictors of student retention usinglistwise deletion as the preferred missing data technique?
    b. Which of the above variables are the best predictors of student retention usingmean substitution as the preferred missing data technique?
    c. Which of the above variables are the best predictors of student retention usingmaximum likelihood as the preferred missing data technique?

2. How does the overall prediction model change depending on how missing data are inputted?

This study began with a literature review that situated the research in a critical paradigm by employing two theoretical frameworks: critical data studies (Iliadis & Russo, 2016) and quantitative critical race theory (Gillborn et al., 2018). Subsequently, the review then paralleled two stories: evolution of power through big data and the utilization of predictive algorithm in higher education to increase student retention. The stories converged as literature around data manipulation was explored which led to chapter three on current and prospective methods of exploring and imputing missing data through various statistical approaches. This quantitative study employed a cross-sectional and longitudinal research design used to predict the retention which served as the outcome variable. The sample consisted of first-year college going students attending the University of Colorado Colorado Springs (UCCS) from Fall 2017 to Spring 2021, which allowed for multiple cohort comparisons across semesters. Data were collected through university enrollment reports, degree conferrals, and University of Colorado Student Integrated Systems (CU-SIS). Analyses were completed using the statistical package, *IBM SPSS Statistics 26.0 for Windows*. Exploratory analysis was first conducted followed by thedevelopment of three models of missing data approaches including listwise, mean substitution, and multiple imputation (described in detail in chapter three). Lastly, a logistic regression was run on each model using all the above independent variables (i.e., demographics, employment, credit load, engagement, high school GPA, SAT/ACT score, current college GPA, major, and financial aid status) to predict the dependent variable (student retention). A logistic regression was used because the dependent variable was dichotomous (i.e., retained or not retained).

Chapter four of the study was designed to interrogate how using each modelof missing data impacted the outcome and statistical results of the dataset including the examination of overall variance and model significance. Chapter five examines the implications of using these techniques on policies, programs, and assessments. Areas for consideration as well as recommendations for student retention data exploration wereincluded.

Assumptions, Limitations, and Delimitations

This study involved a series of assumptions, including that the dataset being used was representative of the student population of the university. It was also assumed that the independent variables that had been selected to predict retention (dependent variable) were the most accurate based on the literature. One of the major delimitations was the use of the specific three missing-data techniques and their level of accuracy for what the research is attempting to prove, which was how each student retention prediction model can change depending on how missing values were inputted. These methods were selected because they not only varied greatly from each other, but they also were widely used in the field of data analytics (Chetverikov, 2019). The limitations of this study were that it was bound to only use values and cases that were missing within the given dataset which was later found that the variable student employment could not be collected due to the lack of accessibility and permission.

Researcher Positionality

As data become more readily available and used to make critical decisions beyond the realm of higher education, there is a need to understand how the information is displayed based on the world view of the researcher (Holmes, 2020). Positionality “reflects the position that the researcher has chosen to adopt within a given research study” (Savin-Baden & Major, 2013, p.

71). This recognizes that all research is situated within a social-political-historical context and consistently influenced by outside factors of the social world (Corlett & Mavin, 2018). The lens through which a researcher views society "impacts interpretation, understanding, and, ultimately, theirbelief in the truthfulness and validity of other's research that they read or are exposed to" (Holmes, 2020, p. 3). This disclosure allows the reader to be more well-informed of the objectivity of the data, results, and implications within the study (Dean et al., 2018; Shaw et al., 2020).

I identify as an able-bodied, cisgendered, White woman who has been privileged to spend the last decade researching meaningful topics across the United States. Since 2010, I have researched parental attachments in the relationships across development (RAD) lab and fed people chocolate cake while exploring impulse control, working memory capacity, and cognitive resource depletion. I have published research on post-traditional student affairs practitioners and have spent countless hours studying the auditory pathways of birds and mice while working in a comparative bioacoustics lab. During my time in the Division of Cognitive and Behavioral Neuroscience lab, I researched electrophysiology in patients with Multiple Sclerosis (MS) and lupus as well as the gender differences in diagnostic sensitively for sleep apnea. I have served as the main point of contact for a National Science Foundation (NSF) grant studying preservice teachers to more recently, researching the ethics of using big data in multiple stakeholder perspectives.

The significance of displaying my academic life history is because what I have done for the past 11 years boils down to numbers. In each of these labs I quantified actions, coded surveys, and data mined ideas in order to put them into *SPSS*, *MATLAB*, *R*, or another computer algorithm to give me an answer—an answer that surely cannot be wrong because p-value stated

significance (after all the confounding variables were considered). However, during the first course in my doctoral program, I learned all researchers have an intent and bias toward what they want to study and report. Around the same time I started my doctoral studies, I went to a presentation given by Dr. Bennet Omalu, a neuropathologist who took on the National Football League (NFL) and discovered that football players' concussions were having severe impacts on their cognitive function. Dr. Omalu stated that knowledge through research is decided by a small committee of research boards like the National Institute of Health (NIH) and the National Science Foundation (NSF) who ultimately decide who gets funded, and therefore, what knowledge will be published and dispensed to the world. Although I know this is an extreme example, I do not think Dr. Omalu is incorrect in his thinking. In our own research, we, the researchers, decide what variables to collect, what to report, and what the implications and future research looks like.

My intent with this dissertation research was to interrogate data manipulation through missing data techniques and, consequently, critically examine the human element that plays a key role from data input to interpretation. I have argued that data can either serve as an equalizer or it can widen the equity gap depending on how it is imputed into a dataset. As data become more readily available and the road to equity becomes more complicated, we need to be able to ask ourselves whose stories are privileged in educational contexts and whose stories are distorted and silenced in the research we publish.

CHAPTER TWO: LITERATURE REVIEW

Most leaders in higher education use quantitative data analytics to not only assess their institution on key performance measures (i.e., retention, persistence, and graduation) but also as a means to justify credibility for evidence-based decision-making processes that lead campus-wide policies and procedures (Sindhi et al., 2019). However, these data analytics are based off statistical methodology that are grounded in assumptions (e.g., normal distribution and homogeneity of variance) and visibility management, (i.e., specific variable collection or data omission) that are controlled by individual data scientists who ultimately decide the execution of these experimental procedures (Flyverbom et al., 2016). This element of human reasoning has led to research in the field of quantitative critical race theory (Quant-Crit) and critical data studies (CDS; Dalton & Thatcher, 2014) who have argued that "numbers are not neutral" (Gillborn et al., 2018, p.158), and therefore, have called for a critical interrogation of quantitative objectivity in data science (Iliadis & Russo, 2016).

There is little research, however, on the empirical evidence that describes how data manipulation occurs at the hands of humans—only that it *does* occur (Kitchin & Lauriault, 2018; Moraes et al., 2019). Researchers have investigated how human labor plays a role in quantitative methodology through the exploration of heuristics and cognitive bias (Kahneman, 2011) as well as the development of tools such as the Implicit Association Test (IAT) that seek to measure factors associated with implicit attitudes and potential prejudices (Greenwald et al., 1998). Yet, the growing complexity of the role human prejudices play in computational analytics has proven difficult to verify due to the reliance on self-reports (Schimmack, 2019) and inconsistencies in instrument validity and reliability (Fiedler & Bluemke, 2005; Fiedler et al., 2006; Rae & Olson,

2018). The aim of this review is to provide the critical exploration of data utility in higher education and investigate how quantitative information can be manipulated and used as a form of power and control which can result in partisan designed procedures, policies, and processes. This background is used to set the stage for the subsequent chapter of methodology that walks the reader through different statistical analysis that alter interpretations and outcomes through various data modification methods.

The conceptual map below (see Figure 2.1) provides a visual of the key areas of focus in the literature review by exploring how the rise of data analytics has been used for the measurement of student retention in higher education. The review then explains the role of human judgment in the development, evaluation, and interpretation of the dataset used to determine retention through a series of statistical analysis. Predetermined data modifications (targeted variable selection or omission), which are used by researchers and data scientists in the development of datasets, can unintentionally manipulate the data which can then exacerbate systemic forms of inequity in higher education by providing deficit-focused interventions to high-risk student populations. This was specifically examined at the University of Colorado Colorado Springs (UCCS), a comprehensive public research university located in Colorado Springs, Colorado. Like all universities, student success is a key goal of UCCS. In the campus 2030 strategic plan, UCCS outlined the importance of increasing retention efforts through integration of student support services, providing a variety of course platform options, and promoting an inclusive and engaging environment for all students (UCCS, 2020). By examining the data behind these efforts, the research study employed more concrete substantive path to understanding specific goals and the metrics used to examine them.

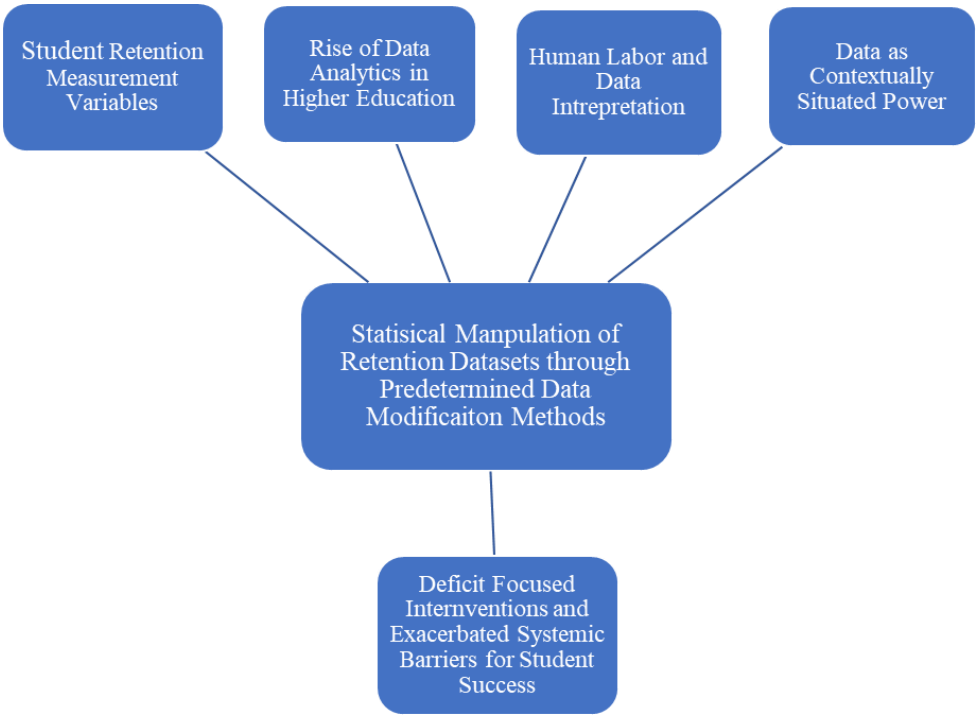


Figure 2.1

Conceptual Map of Literature Review

Theoretical Frameworks

The educational concepts discussed below provide a philosophical foundation that guided the nature of research in the literature review. Beginning with the paradigm to which the subsequent frameworks were situated, this section delves into the ontological and epistemological foundation and researcher positionality for this review. Next, the discussion of the macro-level framework addresses a global academic perspective using quantitative critical race theory (Gillborn et al., 2018); then the discussion focuses on a micro-level perspective by unpacking and reconceptualizing "complex socio-technical systems," (Kitchin & Lauriault, 2018, p. 1) through the nature of critical data studies (Dalton & Thatcher, 2014; Iliadis & Russo, 2016).

Paradigm Lens

This review was positioned from a critical paradigm lens that posits that the ontology (i.e., way of knowing) is a socially constructed phenomena whose purpose is to explore the intersection between power distribution, privilege, and systemic oppression within the context of societal structures (Guba, 1991; Kuhn, 1970). The relevant epistemology (i.e., what counts as knowledge) for this paradigm was situated between the misleading dualistic approach between objective and subjective rationale around the utilization of quantitative data, specifically statistics (Asghar, 2013). In fact, the term statistics comes from the German book *Statistik*, published in 1749, and describes the evaluation of demographic and economic data in relation to the political state of the country (Stigler,1986). Taken together, "the use of statistics is no better or worse than the questions and methods that underlie the research and the social processes in which the research is used" (Guba, 1991, p. 57). This evokes the interplay between the personal values of the researcher and information collected through their scientific exploration. For example, Pearson and Fisher, prominent statisticians in the field of modern mathematics, were both major proponents of eugenics and used their work to propagate their own beliefs about "defending the essentials of the imperial race" (Pearson, 1909, p. 41). In this essence, to be critical (a term used throughout this review) defines "truth" and "interpretation" (Heshusius, 1994, p. 15) as not mutually exclusive, and therefore, must be carefully considered in order to emancipate the societal structures of oppressive regimes that are embedded in everyday culture (Poster, 2019; Joselson, 2016).

Quantitative Critical Race Theory

Quantitative critical race theory (Gillborn et al., 2018), which was adapted from critical race theory (Delgado & Stefancic, 2017), focuses on misrepresentations of quantitative data

that are deeply rooted in institutional processes that result in systemic inequities perpetuated in educational research (Gillborn et al., 2018). There are five major tenets of this theory (Gillborn et al., 2018):

1. Centrality in racism: The pervasiveness and complexity of racism is embedded into the systems of society which makes it difficult to measure and seldom obvious in statistical analysis.
2. Numbers are not neutral: Data collection reflects the researchers' interests and assumptions and is often the result of a majoritarian perception. Therefore, all data should be critically analyzed to understand how and why specific variables were selected and for what purpose.
3. Categories are socially constructed: Labeling variables such as race and ethnicity signal pre-existing qualities that are historically situated which can create patterns that disguise inequities and systems of oppression.
4. Voice and insight: Data do not speak for themselves, and statistics are often open to interpretability and misrepresentation. By controlling for specific variables (for example, socioeconomic status and parental education), statisticians use statistical models to treat social injustices as independent factors which erases the racist influences that are created by inequitable societies.
5. Creating social justice with numbers: The movement toward using quantitative analysis as anti-oppressive praxis and challenge the social construction of the dominant narrative.

Critical Data Studies

Critical data studies (CDS) explore the intersection of the political, cultural, and ethical challenges within the context of big data (Iliadis & Russo, 2016). Originated from the work of Dalton and Thatcher (2014), CDS engages "directly with the cultural regimes of production and interpretation to restore the thick, rich fullness of description that reveals subjects' understandings and intent" (p. 4). It calls attention to the imperfections of algorithm analysis by unmasking the inequities created through the historical variability in uneven collection, evaluation, and application of information. CDS indicates the difference between how data are generated and leveraged, and it specifies the systems to which data are produced and historically situated (Dalton & Thatcher, 2014).

The concept of data assemblages is one of CDS's major components. The principal idea is that data do not exist in a mutually exclusive way, but instead, they are part of an interconnected web of information "bound together in a set of contingent, relational and contextual discursive and material practices and relations" (Kitchin, 2014, p. 23). Data assemblages are comprised of two elements: the apparatus and the element. Both the apparatus and the element are used for categorization and are consistently evolving as markets, technologies, and organizations of change. The purpose of the data assemblage is to show how data is developed by its infrastructures that are intertwined and embedded into the lifeblood of society. This concept explores the notion that data cannot be segregated from the historical context to which it is created and used. The human element in data generation and accumulation is rooted in preconceived judgments and bias that exist both implicitly and explicitly in everyday social contexts. Kitchin (2014) explained how finance data are generated through business models, investments, venture capital, and grants. However, if one were to collect data on the

most successful business models, there would be discrepancy in information. Research has shown significant disparities in the financial sector (and more broadly wealth and income) for marginalized populations which has been a driving factor of economic inequality (Altunbaş & Thornton, 2020; Van Velthoven et al., 2019). Therefore, to gain a more worldview understanding of how finance data is generated, there must be an understanding of greater society and the environment to which the data were amassed (Roberts & Kwon, 2017).

Review of Literature

This section includes a critical analysis of the literature beginning with the current uses of data utilization and transparency in higher education as well as the implications and consequences of data manipulation of student retention data. The literature review included the following search terms: "big data algorithms," "critical studies," "higher education," "retention" AND "missing data" in the following databases: *EBSCO*, *PsycINFO*, *ProQuest*, and *Academic Search Premier*.

Rise of Data Analytics in Higher Education

As technology advances, the capacity to accumulate and store more data increases (Kaisler et al., 2013). Data are now being utilized to develop a new generation of exploration to customize how consumers and businesses view and market products and to reshape how knowledge is viewed and disseminated (Boyd & Crawford, 2012). Data are "now available faster, have greater coverage and scope, and include new types of observations and measurements that previously were not available" (United States Office of the President, 2014, p. 2). Over six years ago, there was approximately four zettabytes of data being generated world-wide. To put this into perspective, one zettabyte can be imagined as every person in the United States of America and Canada taking one photo every second of every day for one month

(United States Office of the President, 2014). Today there are over 40 zettabytes of data being produced, consumed, and analyzed in the world, and this number is rapidly increasing (Maestas et al., 2020). Data advocates argue that quantifiable methodologies provide transparency, replicability, and objectivity to the "age-old search for causality" (Mayer-Schönberger & Cukier, 2013, p. 1143) and legitimizes evidence-based policy making (Sindhi et al., 2019). Modern day computing mechanisms such as machine learning algorithms that use analytics and data mining techniques have revolutionized fields from healthcare (Chen et al., 2017) and marketing (Frizzo-Barker et al., 2016) to agriculture (Carolan, 2017) and medicine (Martin-Sanchez & Verspoor, 2014) by solving complex issues through cost-efficient process optimization.

In higher education, data analytics provide information for colleges and universities regarding student enrollment, retention, and persistence which serve as an important indicator of overall student satisfaction and as a key performance measurement of a university (Rajuladevi, 2018; Viberg et al., 2018). Bichsel (2012) surveyed a sample of members from EDUCAUSE (n = 231), a nonprofit association in higher education, and members of the Association of Institutional Research (n = 135) and showed that 69% of respondents found data analytics to be a major priority for at least some of their departments, programs, or units, and furthermore, that this would rise to 86% in the next two years.

In higher education, research has indicated that using big data can reduce student attrition by providing proactive evidence-based interventions through early-alert systems and by monitoring student engagement in real time (Braxton et al., 1997; Rajuladevi, 2018). Big data advocates argue that by using analytics driven by machines, numbers can "speak for themselves" (p. 8) because they can find patterns that humans may not even be able to detect (Anderson, 2008). For example, Georgia State University has used more than 10 years of student data, over

150 thousand student records, and 2.5 million grades to develop a system with more than 800 alerts that track the daily activity of all undergraduates to identify at-risk behaviors (i.e., grades and attendance), and they have advisers proactively respond to alerts by reaching out in a timely manner to get students back on track (McMurtrie, 2018). Their results have graduated an additional 1,700 student per year, decreased time to graduation by a semester, and created more than 50,000 face-time interactions between students and advisors.

Another example of big data utilization is out of the University of Arizona where researchers have been using "digital traces" to track spatial-temporal aspects of student life through their student identification cards (Blue, 2018). When a student swipes their student ID (CatCard) into the recreation center, the library, dining hall, or any other place that requires CatCard access, they leave a "digital footprint" as to where they have been, how long, and even going as far as to assume who they are hanging out with. For example, if two students frequently check-in to multiple areas at the same time, administration/researchers can assume they run in the same social circle. Blue (2018) showed that by using these traffic patterns, researchers are better able to predict student retention than end of year grades because these patterns start day one; whereas, final grades are posted at the end of the semester. A student may begin to feel overwhelmed by college and seclude themselves, which can be tracked over the semester and used as a form of intervention to assist the student in overcoming any issues they are facing. By using these social integration measures, researchers at the University of Arizona can identify 85% to 90% of students who are at risk of dropping out (Blue, 2018).

Student Retention Measurement Variables

The above cases are examples of how higher education institutions have been working to increase student retention over the last decade. Retention is defined as students who are

continually enrolled within the same institution from fall in their first year to the fall in their second year (National Student Clearinghouse Research Center, 2018). The traditional profile of an undergraduate student is someone who is 18- to 22-years-old, White, residential (living on campus), and attending school full-time (over 15 credit hours per semester; Astin, 1984; Pascarella & Terenzini, 1998; Seidman, 2005). The concern is that these attributes, which have been used as the foundational characteristics for student success for the past 50 years, are still seen as the cornerstone variables that yield the highest prediction of student retention even though institutions are becoming more diverse and student demographics are changing (Reason, 2003). Meaning, although the current higher education landscape is evolving to include more students from diverse backgrounds, those who do not fall into one or all of the above categories are considered to be at a higher risk of attrition (DeWitz et al., 2009). Measuring outdated variables has also been seen in state and federal reporting systems such as the Integrated Postsecondary Education Data System (IPEDS) which state retention is based on "data from first-time, full-time freshman students who graduate within six years of their initial enrollment date (Yu et al., 2010, p. 308). Students who do not fit this criterion (e.g., attending part-time) are not accounted for in the data, and therefore, their characteristics are not evaluated for proactive interventions and policies (Fike & Fike, 2008). Although IPEDS has expanded their inclusion of retention criteria to encompass more inclusive student characteristics, many research studies often still use traditionally studied variables which fail to account for the rapidly changing demographics of the undergraduate student population (Barbera et al., 2020; Delen, 2010; Kai et al., 2017; Ortiz-Lozano et al., 2020).

Recent literature that connects quantitative methodologies and critical studies have begun to include a wider variety of variables such as high school grade point average, first-year college

grade point average, socioeconomic status, race, ethnicity, gender, credit load, financial status, employment, and campus engagement as well as entrance exam scores to predict and understand patterns of student success (Gillborn, 2018; Pérez Huber et al., 2018). Variables such as financial status, employment, and credit load provide more insight into a student's background and move beyond traditional standards of measurement to explain educational outcomes. This holistic metric of understanding retention is critical information given the retention discrepancy between different student populations. According to a 2018 study from the National Student Clearinghouse Research Center, a national research center for postsecondary institutions, "Of the students who started college in fall 2016, 73.9 percent were retained at any U.S. institution in fall 2017, while 61.6 percent were retained at their starting institution" (p. 1). According to the same study, students of color had "the lowest retention rate (67.0%): just over half returned to the starting institution (52.5%) and an additional 14.5 percent returned to an institution other than the starting institution" (p. 2). Further, non-traditionally aged students (over the age of 24) had only a 52.6% retention rate (National Student Clearinghouse Research Center, 2018).

In sum, the move from standardized homogenous measurement of student success to more inclusive variables that provide an intersectional lens of understanding a student's background and narrative is an imperative part of proactive retention measures. However, the sheer volume of data on students being too vast to be analyzed within one model requires researchers to select specific variables (like those stated above) to include in the dataset and analysis (Aljohani, 2016; Rizkallah & Seitz, 2017). Although intentions may be to provide better clarity of student success, the human element in variable selection may result in biases and lead to barriers of opportunity (Boyd & Crawford, 2012).

Data as Contextually Situated Power

Data can serve as an equalizer or it can widen the equity gap depending on how it is utilized (Pitcan, 2016). Although frequently viewed as neutral self-evident units of information that reflect transparent objective facts, data are a form of power that do not merely exist "in the raw" (Gitelman, 2013, p. 2) as a kind of pure evidence. They are "generated" (Gitelman, 2013, p. 2) by researchers and organizations who extract user information through data processing tools that are "hijacked to serve agendas that benefit research and industry" (Iliadis & Russo, 2016, p. 1). These "cooked" (Gitelman, 2013, p. 2), or processed data, can be weaponized and used to elicit emotional responses which can influence decisions and judgments through falsified data "propaganda" (Levy & Johns, 2016, p. 12). This can be seen in the 2020 presidential election, global pandemic, and nation-wide protests over police brutality (King, 2020). Researchers Kitchin and Lauriault (2018) stated, "It is only the uses of data that are political, not the data themselves" (p. 5).

Whether statistics surrounding current events are selected deliberately or unintentionally, data manipulation is ever present in all fields, specifically academia (Levy & Johns, 2016; Starkweather & Herrington, 2018). In a 2015 study by the Open Science Foundation, researchers reproduced over 100 psychology experiments and found two-thirds could not be replicated. This lack of methodological fidelity not only compromised the precision of the statistical analysis but also revealed flaws in "reproducibility which is the hallmark of credible scientific evidence" (p. 943). This is evident when research values innovation more than confirmation, which was seen throughout this review. Further, a recent systematic literature review by Moraes et al. (2019), who examined the social, cultural, and political inequities in computational mechanisms, found

that out of the 15 studies analyzed, over 80% did not discuss equitable practices in data mining and machine learning techniques.

Though datafication can disguise inequities and systems of oppression, so can its absence: "A lack of data is another indication of power, the power to remain hidden" (Iliadis & Russo, 2016, p. 1). Flyverbom et al. (2016) introduced the term "visibility management" (p. 98) which is the level of salience to describe how digital technologies are seen, known, and regulated and the interplay between knowledge and power. The ability to make data visible (i.e., specific variable collection or data omission) involves a form of control around the transparency, disclosure, and accountability of information that are deeply dependent on the acting individual or organization. For example, in the article, "Why Most Published Research Findings Are False," Ioannidis (2005) stated that factors such as (a) bias (e.g., in study design, data analysis, and output interpretation); (b) numerous independent research teams (e.g., the probability of having significant findings of the same research question that is studied by many researchers increasing the chances of finding a false positive); (c) corollaries such as the size of the scientific field (i.e., the smaller the field, the higher probability of having a statistically significant outcome); and (e) design flexibility (i.e., using one-tailed vs. two-tailed hypothesis testing). Ioannidis also stated that financial and other interests also impact what Flyverbom et al. called the "visibility management" (p. 98) of data.

One example of this bias comes from Dr. Omalu, a Nigerian-American physician neuropathologist, who challenged the National Football League and discovered chronic traumatic encephalopathy (CTE; Omalu et al., 2005). At the 2016 Significant Speaker Series presentation at the University of Colorado Colorado Springs, Omalu gave a speech about how major funding organizations such as the National Institute of Health (NIH) and the National

Science Foundation (NSF) ultimately decide what knowledge they put into the world based on allocation of funds (Omalu, Significant Speaker Series, 2016). Omalu indicated that funding is based off a panel of researchers who score applications based on specific criteria, their own research interests, and personal judgments (National Institution of Health, 2020; Omalu, Significant Speaker Series, 2016). Further, once the studies are concluded, the findings are then reviewed by researchers who originally funded their work to ensure criteria was met (National Institution of Health, 2020).

The research panel's level of editability "extends the level of control over the presentation of information means that individuals are often strategic in what, how, and where they present information" (Flyverbom et al., 2016, p. 100). These principles of gatekeeping decide which projects are given the opportunity to come to fruition that will result in publications and dispersion of knowledge into the field as well as the public, governmental, and business sectors. Then the information and data are leveraged as an objective truth used to make arguments and public policies (Stone, 2013). The dynamics of information production through accessibility and transmission boil down the need for critical interrogation of research methods, specifically data disclosure and transparency. Using this framework in the context of higher education, the need to understand stakeholder perspective becomes apparent when seeking to find specific measurements such as student retention.

The Role Human Labor in Data Interpretation

The human element plays a key role from input to interpretation and can influence the assumption that from data derives objectivity because although data do not discriminate, people still do (Haselton et al., 2015). For example, teacher recommendations are often used as a weighted variable when predicting student success and college admission. Previous research has

shown there is significant bias on student performance ratings from teachers based on gender and race (Dee, 2004; Lavy & Sand, 2015). This may skew scores of applicants of different demographic backgrounds and decrease their chances of college acceptance (Capers et al., 2017). Additionally, colleges may collect information on the location of the high school a student attended, but they may not collect data on their performance. Due to the discrepancy between high- versus low-income schools as well as quality of education, potential applicants may be overlooked based on their zip code rather than their actual academic skillset (Bruckner, 2018; Guha et al., 2018).

In a study by Lowry and Macpherson (1988), research on assumed calculation of objectivity was reported through computer analytics at St. George Medical Center. The center began using computer algorithms to screen applicants in their admission process. The researchers found that women and non-European sounding surnames were discriminated against in the initial application screening process. Lowry and Macpherson stated that the computer program was written by a staff member to create a more efficient process and reduce the workload of the admissions committee. Although the researchers did not detail how the computer program arrived at these biases, Lowry and Macpherson stated:

> It is easy to see why women might be discriminated against, [*sic*] there is more risk of them wanting time off work because of family commitments. Likewise, some overseas doctors do not have a sufficient command of English to practice medicine, because understanding the colloquial language is as important as grasping the technical terms. (p. 657)

Lowry and Macpherson (1988) stated discrimination was wrong yet reinforced it in their review of the events. They stated that the program was not presenting new bias but merely reflecting that bias was already in the system. This causes implicit bias associations derived from human judgment during the creation of training data that are used to train the algorithms to work automatically (Williams et al., 2018). This example illustrates how human judgment can

influence how data are entered into algorithms. These judgments are based on beliefs or heuristics which are cognitive tools "that ignore part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods" (Gigerenzer & Gaissmaier, 2011, p. 454; Kahneman, 2011). Tversky and Kahneman (1974) coined three bias heuristics that can influence the way in which data analytics are created and how they are interpreted to various stakeholders: adjustment/anchoring, representativeness, and availability.

In anchoring bias, there is the tendency to heavily rely on one piece of information when making a decision (Schwartz, 2014). Tversky and Kahneman (1974) asked people to name how many African countries were part of the United Nations after they spun a wheel with numbers one to 100. The researchers found that people "anchored" to the number they spun and based their percentage around it. When the wheel landed on 10, they found most participants estimated the number of African countries was around 25%. When the wheel landed on a higher number, the estimate was higher (60; 45%). Even though the wheel spin was arbitrary and there were no correlations between the two, participants anchored themselves to the number they spun and the percentage of countries. Similarly, representativeness heuristic bias is the perception that one datum resembles another even though they may not be correlated. Availability bias states that people will often rely on readily available information rather than viewing a completed data set or discovering the sum of all parts of an analysis. Hence, human biases can impact the objectivity and neutrality of quantitative methodology.

One-way researchers have attempted to measure human biases is through the use of the Implicit Association Test (IAT), a tool that seeks to measure implicit attitudes and potential prejudices (Greenwald et al., 1998). As one of the most influential articles in the field of social psychology, Greenwald et al.'s research on implicit bias has over 4,500 citations and has gained

significant popularity over the last two decades (Schimmack, 2019). The IAT combines the work of Tversky and Kahneman into an instrument that has led to the "examination of unconscious and automatic thought processes among people in different contexts, including employers, police officers, jurors, and voters" (Sleek, 2018, p. 1). The quick synopsis of this test is that participants are given sets of words and images with one being positive and the other being negative. As the words flash in front of the screen, the participants are asked to respond as quickly as they can if they think the words and images are related concepts with the purpose that people will respond faster when ideas are more closely related (https://implicit.harvard.edu/implicit/iatdetails.html).

At first glance, this test seems like a good way for data scientists to see if they are bias in the data they are collecting. By using this tool before data analysis, predisposed prejudices will be exposed and taken into consideration when outputs are interpreted. However, the jury is still out on this instrument, specifically when it comes to construct validity (the degree to which the test measures what it says it is supposed to be measuring; Schimmack, 2019). Payne et al. (2017) argued that although the IAT is intended to measure attitudes, that could be dependent upon situations. Further, they also stated that attitudes are not stable over time, and therefore, should not be used as a fixed measurement of individual differences. Meaning, a data scientist may not have the same attitudes toward specific constructs over time, which proves difficult to argue bias in quantitative data outcomes. Lastly, there is no consensus in the literature about if the IAT is measuring implicit or explicit attitudes (Falk & Heine, 2015; Gawronski & Bodenhausen, 2017; Kurdi et al., 2017; Walker & Schimmack, 2008). Samayoa & Fazio (2017) indicated, "The automatic activation of attitudes should not be equated with individuals' lack of awareness of the attitude" (p. 273). The mixed reviews on the IAT's effectiveness, a widely used tool, demonstrates that measuring an individual's implicit attitudes and potential prejudices with the

intention to argue data's objectivity is a multifaceted and complex case that requires much more clarification.

Critical Implications of Data-Driven Policies, Practices, and Procedures

Taken together, the use of data as power and the influence of human interpretation on these data can have serious implications on policy, regulations, and the level of standards that are used. For example, according to the National Student Clearinghouse Research Center (2018), only 58 % of students who began college after the fall semester of 2012 graduated within six years. Community colleges and for-profit universities had the lowest percentage (below 40%). According to Vice President Voight of Policy Research at the Institute for Higher Education Policy, graduation rates are low because institutions are not adapting to the needs of the students (as cited in Nadworny, 2019). However, since the implementation of the College Scorecard, an initiative by former President Obama that allows students to compare universities regarding financial aid, completion rates, student debt, and other factors that might influence college choice, graduation rates have increased by 1.5% (Nadworny, 2019; U.S. Department of Education, 2019). In a closer examination, the National Student Clearinghouse Research Center showed that students of color (specifically Latinx and Black) had a significantly lower completion rate (41%) compared to other students' demographics (Asian, 71.9%; White, 67.1%). The data from the National Student Clearinghouse Research Center, which uses a proprietary service called "StudentTracker," a matching algorithm, reports enrollment and completion rates of students in higher education institutions. The system matches student demographics (mainly name and date of birth) to their enrollment data (term, status, degree attainment).

In theory, colleges that utilize this system should be receiving rich information about their students; however, in practice, there are major problems with relying on this type of

information. Conaway and Bethune (2015) stated, “Using a first name can elicit a stereotypical perception and suggest that first name stereotypes elicit consequences including an impact on academic achievement” (p. 165). According to Erwin (2006), people who perceived names as more familiar were viewed in a more positive light than those with a unique name. This is also known as the “social-desirability value of the individual’s first name” (Conaway & Bethune, 2015, p. 172). Because the matching algorithm in the National Student Clearinghouse Research Center (2018) used data analytics such as machine learning algorithms, initial inputs may include bias due to name identification (Conaway & Bethune, 2015).

In a cross-sectional case study analysis by Dynarski et al. (2013), which examined enrollment coverage rates of the National Student Clearinghouse Research Center from 2008 to 2010 in Michigan, they found significant sources of missing data and lack of enrollment coverage. One reason for this was because of the Family Educational Rights and Privacy Act (FERPA, 1974), a federal law that protects the privacy of student education records. The law applies to all schools that receive funds under an applicable program of the U.S. Department of Education (U.S. Department of Education, 2007). The term “FERPA blocked” is used to describe student-level information (enrollment and degree information) that cannot be released to research centers. Interestingly, “records of students with different characteristics are blocked at different rates. Over seven percent of Asian/Pacific Islanders’ records were blocked, five percent of Hispanic students’ records, and only about 3 percent of white and black [*sic*] students’ records” (Dynarski et al., 2013, p. 16). Further, the overall percentage of FERPA blocked student records drastically changed from year to year (i.e., 8.5% in fall 2010 to 24.2% in fall 2011; Carnevale et al., 2013; Carnevale et al., 2016). The implications of FERPA blocked records and

mismatching typographical errors in national data centers shows the lack of congruence between intentional reporting and actual data reported.

The use of analytics plays an important part for many fields of study including education. Quantitative methodologies can unify an institution by providing important data about specific student populations that can be cultivated to deliver proactive interventions (Bichsel, 2012; Pilgrim et al., 2017). Research examining how data are collected and reported from national educational research centers and local institutions is vital; however, it is key to focus on the utilization and specific student data and understanding the reasoning for such variables (Mittelstadt et al., 2016; Murumba & Micheni, 2017). At the 2019 National Association for Developmental Education, Joseph Garcia, Chancellor of the Colorado Community College System (CCCS), stated, "We need to focus on outcomes, not enrollment. What are we measuring when we track student success? What are the goals of collecting the data?" As data become more readily available and the road to equity comes more complicated, these critical questions need to be asked.

Conclusion

Throughout this review, the following were established: Firstly, big data have changed the essence of what is believed to be objective knowledge (Boyd & Crawford, 2012). Since its inception, data have been used as a soothsayer by predicting the future and earmarking the past through a "new historical constellation of intelligibility" (Berry, 2011, p. 12). Secondly, we are in an era of information overload. Specifically, higher education institutions are being inundated in data collected from enrollment demographics, learning management systems, co- curricular programming, facility usage, and other aspects of student life within and beyond the walls of the academy (Daniel, 2015; Papamitsiou & Economides, 2014; Picciano, 2012; Williamson, 2018).

Thirdly, numbers are not neutral and can be influenced by social, cultural, and political agendas (Gillborn et al., 2018). Personal values and beliefs are embedded into data collection, variable selection, information omission, output interpretation, and overall transparency (Cox et al., 2014). From data collection methods to the development of the datasets, human agency lends itself to bias ideologies of social control and self-interest (Guba, 1991).

It is critical the recognize that data can structurally change outcomes, specifically in higher education, because of how it is leveraged. Research on implicit bias has not only been researched extensively but deemed disputable (Brownstein et al., 2020; Fazio et al., 1995; Fiedler & Bluemke, 2005). The current study sought to understand data manipulation from a quantitively ontological perspective (i.e., how things exist statistically; Guba, 1991), not a human-centric viewpoint, through the assessment in which the landscape of numbers could be analyzed, formulated, reproduced, and then interpreted. How data are classified will fundamentally change how variables are described, analyzed, and interpreted. In chapter three, I discussed the types of data that can be manipulated, how different statistical analysis can alter interpretations, and outcomes through various data modification methods. This research sought to critically interrogate the notion that "numbers speak for themselves" (Gillborn et al., 2018, p.158) by providing empirical evidence of data manipulation (specifically omission) and asked stakeholders who participate in the generation, aggregation, or dissemination of data to recognize that quantitative information is socially, culturally, and politically situated depending on the individual and their values, which influences how specific data will be accumulated and distributed (Boyd & Crawford, 2012; Williams et al., 2018).

CHAPTER THREE: METHODOLOGY

At the University of Colorado, information regarding admissions, enrollment, student records, campus engagement, demographics, and student financials are stored in a central information warehouse; University of Colorado Student Integrated Systems (CU-SIS). This system is used by various campus stakeholders to report to federal and state entities for accreditation and funding purposes and to provide a vital barometer of institutional effectiveness through the measurement of overall student satisfaction and key performance indicators such as retention, persistence, and graduation (Rajuladevi, 2018). Data extracted from this information warehouse are used to develop evidence-based rationale for decisions regarding policies and procedures at the university-wide level (Sindhi et al., 2019).

However, because data are collected and reported in short "snapshots" (UCCS, 2020b, p. 4) of time (i.e., specifically after fall and spring semester census dates), caution must be exercised when examining accuracy and population representativeness. Reports pulled from datasets that have missing or inaccurate information and are then used to develop policies or campus programs could be missing key characteristics of high-risk student populations that are needed for proactive interventions. Further, statistical analysis of the observed datasets is conducted by humans, who are prone to a variety of bias, may change the results and interpretation of the outcomes (Flyverbom et al., 2016; Yoon et al., 2018). The following research was not intended to target specific campus stakeholders for intentionally manipulating data; rather, it was to hold the mirror up to data users across higher education institutions to show that "numbers are not neutral" (Gillborn et al., 2018, p. 158) and to call for a critical interrogationof quantitative objectivity in data science (Iliadis & Russo, 2016; Shields, 2005).

Research Questions

The research situated missing data approaches (described in detail below) to the connection of higher education by using retention, a key performance variable for first-year students according to the Integrated Postsecondary Education Data System (2020), as the dependent variable for the study to provide real-time examples of how data are used at post-secondary institutions. The following overarching research question directed this manuscript: Does using different methodological techniques for missing values in each dataset impact the interpretation of the statistical outcome? This was addressed by exploring the following sub-questions pertaining to student retention:

1. How differently does the utilization of missing data techniques predict student retention using a combination of student demographics, credit load, campus engagement, high school GPA, standardized test scores (SAT/ACT), current college GPA, major, and financial aid status?
    a. Which of the above variables are the best predictors of student retention using listwise deletion as the preferred missing data technique?
    b. Which of the above variables are the best predictors of student retention using mean substitution as the preferred missing data technique?
    c. Which of the above variables are the best predictors of student retention using maximum likelihood as the preferred missing data technique?
2. How does the overall prediction model change depending on how missing data are inputted?

These specific variables were selected because they have been designated as crucial indicators for predicting student retention (Picciano, 2012; Rajuladevi, 2018; Sander, 2016).

Selection of Methodological Techniques

The following section describes the most common methods for handling missing values within a dataset and how they were utilized within each contextual framework. A description of each method along with assumptions and bias are explained. This section provides the reader with what techniques are most used for missing data and what was selected for the current study.

Methods for Incomplete Data

Methods for handling missing data should be specifically tailored to the dataset of interest (Allison, 2002; Salgado et al., 2016); therefore, dataset and algorithm development are critical. Methods should be chosen based on sources of missingness such as MCAR, MAR, or NMAR because the "mechanisms by which the data are missing will affect some assumptions supporting [the] data imputation methods" (Salgado et al., 2016, p. 13). The most widely used methods of handling missing data fall into three main categories: (a) deletion which consists of eliminating all cases that have missing data (i.e., complete-case analysis, listwise deletion); (b) available-case analysis (pairwise deletion) which consists of single imputation filling in missing data with one method/rule (i.e., mean/mode substitution, linear interpolation, hot-deck/cold-deck); and (c) model-based methods which consists of using predictive algorithms to estimate the missing data (multiple imputation and maximum likelihood; Schafer, 1999). Each of these methods is specifically chosen based on the researchers' or data scientists' (often implicit) standards of justification (Anseel et al., 2010; Cox et al., 2014).

Deletion Methods

The simplest way of handling missing data is to rid the dataset of any missing cases. Excluding all cases with missing data is one method frequently used in statistical software packages such as *SPSS*, *SAS*, and *R* and is found to often be the default method in fields such as

psychology and education (Peugh & Enders, 2004). In general, this method is valid for only MCAR data because assumed missingness is not a function of the outcome or dependent variable (Cox et al., 2014). Missing variables are unrelated to what is being measured in the analysis, so they would likely be less bias (Pepinsky, 2018). The most common deletion methods include complete-case analysis, commonly referred to as listwise deletion, and available-case analysis, also known as pairwise deletion.

During listwise deletion, all cases that have missing variables are deleted from the dataset. For example, in a comparative analysis that explored the use of data mining techniques and student retention, Delen (2010) investigated a total of 39 variables ranging from student demographics, academic performance, and financial status. Listwise deletion requires that any variables missing in any category (i.e., demographics, academic performance, and financial status) would deem that the entire case be discarded. The advantage of this type of method is utilizing a complete dataset that can compare all variables because any missing cases are thrown out; however, disadvantages to this method are substantial, such as greatly reducing the sample size which results in less power in significance tests (Baraldi & Enders, 2010).

If MCAR assumptions are violated and missing variables are in fact related to each other, "the analyses will produce biased estimates" (Baraldi & Enders, 2010, p. 10). In the case of Delen (2010), all variables were related because they were all part of a subset population of students who did not have an equivalency to the variables they were measuring. This is problematic because the reason for missing data may not be random, and therefore, Delen's findings and implications may be inaccurate due to inferences about a total population. Further, Raaijmakers (1999) showed that listwise deletion led to a "reduction in statistical power between

35% (if 10% of the data are missing) and 98% (if 30% of the data are missing)" (as cited in Lodder, 2013, p. 1).

In available-case analysis, also known as pairwise deletion, cases are omitted only if they are going to be used in a specific analysis (Newman & Cottrell, 2015). In the same study, Delen (2010) "removed all international student records from the dataset because they did not contain some of the presumed important retention predictors (e.g., high school GPA, SAT scores)" (p. 501). The basis of these analyses is that subsets of data are removed depending on where values are missing. One advantage to this technique is that it increases power of the analysis because it does not completely remove all cases like that of listwise deletion (Wothke, 1993); however, each analysis may be conducted on a different subset of data that can be confusing to interpret and lead to inaccurate results (Cheema, 2014).

Single Imputation

Another technique for handling missing data is called single imputation which includes mean substitution, regression-based imputation, and matching methods (i.e., hot- and cold-deck imputation). Each of these techniques fills in missing data with new values, and the "imputed values are assumed to be the real values that would have been observed when the data would have been complete" (Eekhout et al., 2012, p. 730). The advantage of this technique is the preservation of the entire sample and a reduction of variability in the data; however, no imputation method can provide an exact value, and therefore, complete accuracy should not be assumed (Baraldi & Enders, 2010).

Mean substitution is the simplest and easiest way to input missing variables (Eekhout et al., 2012). One way to utilize this technique is by replacing the mean of the subgroup to the specific case (i.e., using the mean of all women's SAT scores to replace a missing value for one

woman's SAT score; Cox et al., 2014); although, the simplicity of this technique is also its downfall. Research has shown that mean imputation yields highly biased measures of parameter estimation (Knol et al., 2010; Kwak & Kim, 2017; Zhang, 2016). In the SAT score example, it is assumed that the woman whose data are missing will fall within the parameters of the mean for all women's scores. This is especially problematic if there are a lot of missing values and they are MNAR. Nevertheless, some studies (Harrell, 2015) showed that mean imputation is beneficial if "less than 10% of the data are missing and when the correlations between the variables are low" (Lodder, 2013, p. 3).

Regression-based imputation creates a predictive model that uses non-missing variables to predict the variable that is missing. For example, suppose that some SAT scores are missing in a dataset, and a set of other variables such as high school GPA and attendance records do not contain missing values. This technique can then predict the missing SAT scores values by using the non-missing high school GPA and attendance records as predictors in a regression analysis of SAT score on high school GPA and attendance records. As a result, all missing SAT scores are replaced with the predicted SAT scores. The advantages of this technique are the preservation of a normally distributed dataset. It can also be used when more than 10% of data are missing and if variables are highly correlated (Lodder, 2013). The disadvantage is that it can lead to biases if data are not missing at random, or it may lead to implausible values (i.e., negative SAT scores; Hron et al., 2010).

The last single imputation technique is called matching methods or hot- and cold-deck imputation. This involves replacing missing data points with values from another "matched" (Cox et al., 2014, p. 8) case within the dataset (hot-deck) or a dataset that is similar (cold-deck). One example of a hot-deck imputation would be if a 22-year-old White male (Student A) did not

report his parent's income in a survey, then the missing value (parent's income) would be inputted from a similar participant (Student B) in the dataset who did report their parent's income. A cold-deck imputation would use Student A's parent income from an earlier dataset (of when he was 20). The advantage of this technique is that realistic values and plausible values are used for missing data; however, this also creates implicit assumptions and removes random variation in the sample (Chhabra et al., 2017; Cox et al., 2014).

Model-Based Methods

The last category of commonly used methods for dealing with missing data is model-based methods which consists of using predictive algorithms to estimate the missing data (Madley-Dowd et al., 2019). Methods such as multiple imputation and maximum likelihood are considered more modern approaches because they are more robust and missing values are not replaced or inputted; rather, used as an "estimate for each missing value through simulated values but rather to represent a random sample of the missing values" (Yuan, 2010, p. 1).

Pilot Study

To demonstrate the impact of missing data analyses, a pilot study was conducted using a dataset from Colorado State University's Center for the Analytics of Learning and Teaching which centers on using learning analytics to inform educators, administrators, and students about learning behaviors. In this study, a dataset on self-regulated learning and motivation was analyzed. Data were collected on previous academic achievement scores (i.e., high school GPA, ACT Math scores), self-regulation using the motivated strategies and learning questionnaire (MSLQ) explained in detail below (Pintrich & DeGroot, 1990), and MATH 160 (Calculus 1) exam scores. This dataset provided real-life information regarding sources of data missingness and corresponding approaches to handling incomplete datasets.

To thoroughly understand potential data manipulation, there must be a careful consideration in the relationship between the nature of the chosen statistical analysis, the reasons for missing data, and the missing data patterns. This pilot investigation provided cumulating evidence of this triaged approach that was used for later research methods. In the following sections, a description of the dataset (i.e., demographics, measures, procedure) was provided along with an examination of missing variables to determine level of randomness. Depending on how variables were missing (i.e., source of missingness), missing data were imputed based on the above methods (i.e., deletion, single-imputation, and model-based) and then ran through a series of analysis. The purpose of this pilot study was to show how data results change depending on methodological approach to value-missingness.

By using a smaller dataset, the reader can view the impact of data manipulation without the complexities that come with massive amounts of information in larger data repositories. For example, in the field of comparative bioacoustics, many researchers use mice and birds to study auditory processes because of the simplicity of their neural networks. Once there is an understanding of the basic principles of behavior and anatomy, research is then conducted on more complex beings like humans and primates (Hopp et al., 2012). This level of building on complexity is the most fundamental method to understanding intricate phenomena (Bland, 2015).

Participants

Students enrolled in first semester calculus at Colorado State University were recruited to participate in the study. Out of approximately 460 students who registered for calculus in the Spring semester of 2016, a total of 415 students agreed to participate in the study. Of the 415 participants, 200 majored in engineering, 11 in mathematics, 42 in computer science, three in statistics, and five in physics. Students identified as Caucasian were 279, 48 Hispanic, seven

Asian, 25 international, and 33 others. Most of the participants were first- and second-year students with 64.8% freshman, 20.4% sophomore, 4.5% juniors, and 0.7% seniors. In addition, 20.9% of the participants were identified as first-generation university students.

Measures

The MSLQ (Pintrich & DeGroot, 1990) was used, a self-report instrument that utilizes a 7-point Likert scale (1 = not at all true of me and 7 = very true of me) designed to measure students' motivation and self-regulated learning (SRL). The subscales of MSLQ that related to SRL and motivation were administered which consisted of critical thinking, metacognition, effort regulation, self-efficacy, and time and study environment. Critical thinking referred to the degree to which students reported applying previous knowledge to new situations to solve problems, reach decisions, or make critical evaluations. Metacognition referred to the awareness, knowledge, and control of cognition. Effort regulation referred the students' ability to control their effort and attention in the face of distractions and uninteresting tasks. Self-efficacy referred to judgments about one's ability to accomplish a task as well as one's confidence in one's skills to perform that task. Time and study management refers to scheduling, planning, and managing one's study time. This included not only setting aside blocks of time to study, but the effective use of that study time and the setting of realistic goals (Pintrich et al., 1991).

Procedure

During the first week of class, students were encouraged to fill out the MSLQ. There was a total of three exams and a final. However, at the time of the data collection, only exam 1 scores were available, and this was used as a measure of achievement in calculus learning.

Pilot Research Questions

The general research question (Question 1) served as the overarching guide in which this pilot was situated. Question 1a.-c. was specific to the pilot dataset and served to answer the sub-questions (listed below) by employing three different methods for missing value imputation:

1. Does using different methodological techniques for missing values in a given dataset impact the interpretation of the statistical outcome?
    a. How well does the combination of ACT MATH scores, prior High School GPA, MSLQ scores, gender, year in college, and type of Major (STEM/Non-STEM) predict Exam1 score using listwise deletion, mean substitution, or maximum likelihood approaches for missing variables?
    b. Does the prediction combination change depending on how missing data are inputted?

Analysis of Pilot Study

All data analyses were completed using the statistical package, *IBM SPSS Statistics 26.0* for Windows. The summary of missing values (Figure 3.1) shows the variables, cases, and values of the observed dataset. The first pie chart stated that out of the 12 variables put in the analysis, all of them had some missing data. The second pie chart stated that out of the 415 total cases, 85 (20.48%) were missing. The last pie chart, which examined all of the values in the dataset, showed that 546 (10.96%) were missing.

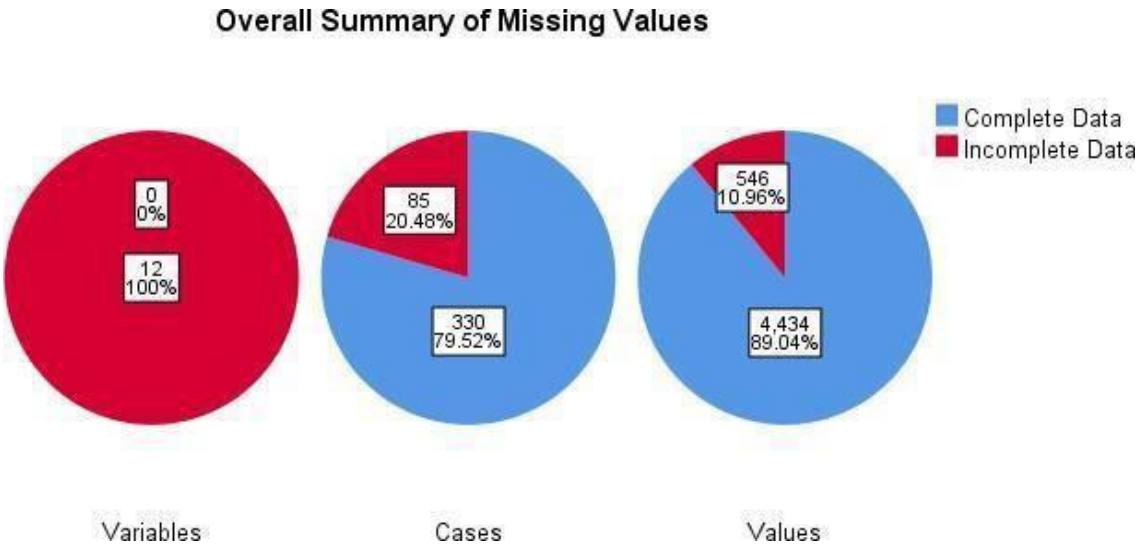


Figure 3.1

Summary of Missing Values

Next, a complete cases analysis was performed (Table 3.1) to provide a breakdown of the percentage of cases used for each variable. This analysis showed that MSLQ variables had the same amount of missing data (77 excluded), and major, gender ACT score, and high school GPA all had the same number of missing data (26 excluded). Next there was an examination to if determine patterns of missing data were related. This was critical to determine what statistical test should be used for the final analysis and missing data imputation. Although the only true way to distinguish between NMAR and MAR would be to contact the participants individually (Leavitt,2019), a missing values analysis indicated that Little's (1988) test of missing completely at random (MCAR; Table 3.1) was not significant, $\chi 2$ 35.545, $df = 19$, $p = .012$; meaning, there was no evidence to suggest that the data were not MCAR.

Table 3.1

Case Summary Analysis

	Cases included		Cases excluded		Total	
	N	Percent	*N*	Percent	*N*	Percent
Exam one scores	410	98.8%	5	1.2%	415	100.0%
High school GPA	389	93.7%	26	6.3%	415	100.0%
ACT math	389	93.7%	26	6.3%	415	100.0%
Student class	389	93.7%	26	6.3%	415	100.0%
Self-Efficacy (MSLQ)	338	81.4%	77	18.6%	415	100.0%
Critical thinking (MSLQ)	338	81.4%	77	18.6%	415	100.0%
Metacognition (MSLQ)	338	81.4%	77	18.6%	415	100.0%
Effort regulation (MSLQ)	338	81.4%	77	18.6%	415	100.0%
Time & study environment (MSLQ)	338	81.4%	77	18.6%	415	100.0%
Major (STEM/Non-STEM)	389	93.7%	26	6.3%	415	100.0%
Gender	389	93.7%	26	6.3%	415	100.0%

Table 3.2

Little's Test of MCAR

Estimated Marginal Means							
Exam score	High school GPA	ACT math	Self-Efficacy (MSLQ)	Critical thinking (MSLQ)	Metacognition (MSLQ)	Effort regulation (MSLQ)	Time & study environ-ment (MSLQ)
71.79	3.43	21.13	5.20	4.158	4.68	4.97	5.60

Note. Little's MCAR test: Chi-Square = 35.545, *df* = 19, Sig. = .012

Statistical Analysis

In this section, research question 1a is addressed using a multiple regression analysis as well as a discussion of the differences in results and interpretation:

1a. How well does the combination of ACT MATH scores, prior High School GPA, MSLQ scores, gender, year in college, and type of Major (STEM/Non-STEM) predict Exam1 score using listwise deletion, mean substitution, or maximum likelihood approaches for missing variables?

Listwise Deletion

Multiple regression was conducted to investigate the best prediction of Exam 1 scores. The descriptive of the analysis can be found in Table 3.3. A total of 330 cases were used in the model. The combination of variable to predict exam one scores was statistically significant, $F(11,318) = 3.18$, $p < .001$. The beta coefficients are presented in Table 3.4. Note that major (if a student was STEM or non-STEM), Self-efficacy (MSLQ) significantly predicted exam 1 scores when all variables were included. The adjusted R^2 value was 0.086. This indicates that almost 9% of the variance in exam 1 scores was explained by the model.

Table 3.3

Descriptive Statistics (Listwise)

	Mean	*SD*	*N*
Exam score	72.81	14.67	330
Ethnicity	1.73	1.47	330
Student class	1.4	.72	330
Gender	.68	.465	330
High school GPA	3.49	1.07	330
Major	.78	.41	330
ACT math	21.29	11.37	330
Self-Efficacy MSLQ	5.22	.94	330
Critical thinking MSLQ	4.14	1.01	330
Meta cognition MSLQ	4.67	.70	330
Effort regulation MSLQ	4.97	.61	330
Time & study environment (MSLQ)	5.60	.79	330

Table 3.4

Summary of Coefficients

	Unstandardized coefficients		Standardized coefficients		
Model	*B*	Std. Error	Beta	*t*	Sig.
Constant	55.702	8.734		6.377	.000

Ethnicity	-.816	.539	-.082	-1.515	.131
Student class	1.527	1.281	.075	1.192	.234
High school GPA	1.173	1.022	.086	1.147	.252
Major	6.009	2.078	.170	2.892	.004
ACT math	.017	.079	.013	.219	.827
Self-Efficacy MSLQ	3.504	.971	.224	3.608	.000
Critical thinking MSLQ	1.565	.978	.117	1.600	.111
Meta-Cognition MSLQ	-4.470	1.773	-.214	-2.521	.012
Effort regulation MSLQ	-.916	1.604	-.038	-.571	.568
Time & study environment (MSLQ)	1.433	1.279	.078	1.121	.263
Gender	-.180	1.750	-.006	-.103	.918

Note. Dependent variable: Exam1NoZeros.

Mean Substitution

Multiple regression again was conducted to investigate the best prediction of exam 1 scores. The descriptive of the analysis can be found in Table 3.5. A total of 415 cases were used in the model. The combination of variables to predict exam one scores was statistically significant, F (11,403) = 5.57, p <.001. The beta coefficients are presented in Table 3.6. The same variables (major and self-efficacy) significantly predicted exam 1 scores when all variables were included. The adjusted R^2 value was 0.108. This indicates that almost 11% of the variance in exam 1 scores was explained by the model.

Table 3.5

Descriptive Statistics (Mean Substitution)

	Mean	*SD*	*N*
Exam Score	71.77	15.60	415
Ethnicity	1.77	1.45	415
Student Class	1.39	.68	415
High School GPA	3.43	1.05	415
Major	.76	.41	415
ACT Math	21.19	11.19	415
Self-Efficacy MSLQ	5.21	.84	415
Critical Thinking MSLQ	4.15	.98	415
Meta-Cognition MSLQ	4.68	.63	415
Effort Regulation MSLQ	4.97	.56	415
Time/Study Environment (MSLQ)	5.60	.72	415
Gender	.71	.440	415

Table 3.6

Summary of Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
Model	*B*	Std. Error	Beta	*t*	Sig.
(Constant)	50.280	8.748		5.747	.000
Ethnicity	-1.135	.512	-.105	-2.217	.027

Student class	2.082	1.238	.091	1.681	.094
High school GPA	2.136	.917	.145	2.329	.020
Major	6.210	1.918	.165	3.238	.001
Math ACT	.081	.074	.058	1.097	.273
Self-Efficacy MSLQ	3.499	1.002	.190	3.491	.001
Critical thinking MSLQ	1.790	1.007	.113	1.778	.076
Meta-Cognition MSLQ	-4.298	1.814	-.175	-2.369	.018
Effort regulation MSLQ	-1.213	1.634	-.044	-.742	.458
Time & study environment (MSLQ)	1.203	1.308	.056	.920	.358
Gender	.796	1.736	.022	.459	.647

Note. Dependent variable: Exam1NoZeros.

Multiple Imputation

Multiple regression was used to investigate the best prediction of Exam 1 score for the final analysis. However, using multiple imputation required a series of steps before the regression analysis could be completed. This technique predicts missing data by using various iterations of data methods until it produces a "best fit" (Cox et al., 2013, p. 382) with the values that are already present. To complete these iterations, a Mersenne Twister was used to set the random number generator and then imputed missing data values. Selected variables were then used in the model. The analysis maintained the default of five imputations (meaning the model was simulated five times and then averaged to predict the missing value). After the multiple imputation was completed, multiple regression analysis was conducted. The descriptive of the analysis can be found in Table 3.7. A total of 415 cases were used in the model. Using the fifth

iteration of variables to predict exam one scores, the model was statistically significant, $F(9,405) = 8.16$, $p < .001$. The beta coefficients are presented in Table 3.8. Major, self-efficacy, and meta cognition significantly predicted exam 1 scores when all variables were included. The adjusted R^2 value was 0.137. This indicated that almost 14% of the variance in exam 1 scores was explained by the model.

Table 3.7

Descriptive Statistics (Mean Imputation)

	Mean	*SD*	*N*
Exam score	71.80	15.62	415
Student class	1.44	0.68	415
High school GPA	3.35	1.09	415
Major	0.74	0.44	415
Math ACT	20.11	12.47	415
Self-Efficacy MSLQ	5.23	0.94	415
Critical thinking MSLQ	4.17	1.07	415
Meta-Cognition MSLQ	4.68	0.69	415
Effort regulation MSLQ	4.96	0.62	415
Time & study environment (MSLQ)	5.59	0.79	415
Ethnicity	1.88	1.40	415
Gender	0.70	0.46	415

Table 3.8

Summary of Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	*t*	Sig.
(Constant)	47.0	9.05		5.19	.000
Student class	2.71	1.40	0.11	1.93	.061
High school GPA	2.28	.88	0.16	2.57	.010
Major	5.71	1.90	0.17	3.00	.003
Math ACT	0.06	.077	0.05	.83	.407
Self-Efficacy MSLQ	3.58	1.10	0.19	3.25	.003
Critical thinking MSLQ	2.05	.91	0.14	2.24	.025
Meta-Cognition MSLQ	-4.74	1.75	-0.22	-2.70	.008
Effort regulation MSLQ	-.83	1.62	-0.03	-0.51	.609
Ethnicity	1.10	.60	-.11	-1.82	.078
Gender	.65	1.86	0.04	0.35	.726
Time & study environment (MSLQ)	1.49	1.22	0.08	1.21	.224

Conclusion and Proposed Research

Each of the above analyses used the exact same dataset and statistics and came out with different results based on how missing data were inputted into the dataset. For example, in the multiple imputation model, ethnicity, student class, critical thinking, and high school GPA were marginally significant compared to the listwise deletion methods where they were not close to being statistically significant. Further, the percentage of variance nearly doubled depending on the missing data approach. This dataset had approximately 10% of the data missing and showed a simplified version of how a data scientist/researcher can change the information, results, and interpretation depending on their chosen techniques. For instance, as a researcher using the listwise deletion technique, one would conclude that a student's major (specifically if they are STEM or non-STEM) is related to how well they will do on their Calculus 1 exam. If a researcher were using a mean substitution as the method of choice, the focus might be on the student's major but also on their meta-cognition and self-efficacy. If there was a need to conduct a more modern approach to missing data, using a multiple imputation method to conduct computational simulations would result in observing that ethnicity, student class, critical thinking skills, and high school GPA coupled with major, metacognition, and critical thinking all can impact how well a student does on their calculus exam. Further, these data also need to be used with caution because missingness is not random. Therefore, if a student decided not to fill out the MSLQ they might be completely removed from the study. There may be a commonality between those who did not complete the MSLQ and those who did; however, this would not be known if only those who had completed cases were counted and used in the analysis.

Dissertation Research Design

The following section describes in detail how the proposed study of predicting student retention using various missing data techniques was conducted. This includes a deeper understanding of the sample as well as the data analysis plan. This study was conducted using an ex post facto, cross-sectional as well as longitudinal research design. This was chosen because the data was collected retrospectively without interference from the researcher. I examined students' characteristics across many variables in at one point in time (enrolling in UCCS as first-time undergraduates) and then reexamined later to measure if they were retained. Measuring retention is often conducted using a retrospective or ex facto methods due to the need for pattern examination over consecutive semesters (Millea et al., 2018). Importantly, the goal of the research was not to predict retention, per se, but to examine how the model that predicts retention changed using missing data methodologies.

Participants and Site

The sample consisted of first-year college going students attending the University of Colorado Colorado Springs (UCCS) a comprehensive public research university located in Colorado Springs, Colorado. As of fall 2019, UCCS has a total student enrollment of 12, 197 of which 84% were undergraduate students (N = 10, 246) who were the target sample population. This site was selected because of sponsor access, feasibility, and the significant amount of data missing from student retention categories. For example, in the Demographic Information Fall 2019 Cohort from First-Year Cohort Retention Report (UCCS, 2020), a total of 1,787 reported their gender, race, or ethnicity but only 1,391 reported their estimated income. Additionally, 1,788 students reported their high school GPA; whereas over 1,250 did not report their ACT scores and over 200 did not report SAT scores. Although this is a snapshot of the factors used to

determine retention, the varying samples were cause for concern because of potential discrepancies in the data.

Data were collected from fall 2017 to spring 2021. This length of time allowed for the collection of retention data, which was used as the dependent variable. Retention is defined as students who are continually enrolled within the same institution from fall in their first year to the fall in their second year (National Student Clearinghouse Research Center, 2018. Using subsequent years allowed for comparison of multiple cohorts. At UCCS during Fall 2018-Fall 2020, the average incoming first-year class was approximately 1,500 students which equated nearly 4,500 students across six semesters. Criteria for participant inclusion comes from the National Student Clearinghouse Research Center (2020) national benchmarks which stated retention is measured using data from students with no previous college enrollment in the four years prior to the entering cohort year (i.e., no transfer students) with degree-seeking status. Both full-time (enrolled in 12 or more credit hours per semester) and part-time (less than 12 credit hours per semester) students were included.

Measures

Specific independent variables have been selected because they have been designated as crucial indicators for predicting student retention (Picciano, 2012; Rajuladevi, 2018; Sander, 2016). These included student demographics (i.e., first generation status, ethnicity, gender, age) to understand overall sample characteristics, employment status (i.e., estimated number of hours working) which examined how often students are working (both on-campus and off-campus). Previous research has concluded that student working full-time are less likely to be retained (Astin, 1984; Pascarella & Terenzini, 1998). However, with the rising cost of tuition and growth in attendance of non-traditional students who are already working in the field, there is an

increase in both student employment (working on campus) and overall employment (Seidman, 2005). At UCCS, nearly 1800 students are working on campus and over 10% are working while attending school (UCCS, 2020). Course load/credit hours taken per semester was measured in this study. The average credit load at UCCS is 12 credits per semester which is three credits lower than most universities. A fall 2018 survey conducted by the UCCS Department of Institutional Research stated that many students were not taking the traditional 15 credit full-time course load because they were working, have family obligations, or are worried about their academic performance. Tinto (1993) stated engagement (i.e., if student attended at least one event on campus or participated in university club or organization) is often used as a predictor of student retention because engagement creates a sense of belonging and helps students develop meaningful connections on campus. Other notable predictors include high school GPA, SAT/ACT score, current college GPA, major, and financial aid status (i.e., use of loans, grants, or scholarships). All variables were previously collected by the Office of Institutional Research for reporting to the IPEDS.

Data Collection

Data were collected through university enrollment reports, degree conferrals, and University of Colorado Student Integrated Systems (CU-SIS). All information was compiled by the Office of Institutional Research and anonymized to ensure student data privacy and federal regulations are compliant. Data records were stored in a password encrypted portal and was only accessed by the researcher.

Data Analysis

All data analyses were completed using the statistical package, *IBM SPSS Statistics 26.0* for Windows. A complete case summary as well as descriptive statistical analyses were

performed on the sample to obtain a clear understanding of the dataset and to indicate how many values, cases, and variable were missing. The missing value analysis were computed to explore the pattern of data missingness along with Little's MCAR test which examined the degree data were missing at random. All missing data were filtered through three different approaches including listwise: (a) excluding all cases with any value missing, mean substitution; (b) replacing missing values of a specific variable with the mean value of the observed (non-missing) specific values; or (c) multiple imputation which uses a predictor to impute variables that have missing data. A logistic regression using all of the above variables (demographics, employment, credit load, engagement, high school GPA, SAT/ACT score, current college GPA, major, financial aid status) was imputed as independent variables and was used to predict the dependent variable (student retention). The logistic regression was used because the dependent variable was dichotomous (retained or not retained) unlike the pilot study which used a multiple regression because the dependent variable was categorical (exam 1 scores).

CHAPTER FOUR: RESULTS

The purpose of this study was to investigate the process of only using an observed subset of data to generate overarching conclusions and generalizations about larger populations, specifically those from underrepresented or marginalized communities. Unlike most clinical trials that can keep a pulse on the design and implementation stages of data collection, the complex data resources and burdensome processes in higher education have made finding and resolving incomplete information more difficult to control or anticipate (Pitcan, 2016).

Because there is no universal method to analyze missing data, many educational researchers use their own judgment when analyzing incomplete data which leaves room for bias, error, and assumptions about the information (Flyverbom et al., 2016). The choice to delete, predict, or substitute missing values is often not studied because it is frequently viewed as secondary analysis rather than the main point of focus (Bichsel, 2012; Moraes et al., 2019). This is significant in the overall understanding of how data are understood and how data are contextualized.

To understand the frequent yet widespread problem of incomplete information, the following chapter provided a dissection of commonly used missing data techniques. To complete the narrative of how missing data are handled, this chapter addressed the acquisition of the data and how the information was screened and cleansed for missingness. Lastly, the chapter closed with the summary analyses of the results.

Data Acquisition and Preliminary Analysis

The data came from the University of Colorado Colorado Springs (UCCS) Office of Institutional research in a password encrypted excel file. The raw data had 28 variables which

included: fake identification (pseudo identification that was used to link the students data), term code (code specific to the learning management system to identify which semester the students were actively enrolled in), term code description (common language used for the term code such as Fall 2017), enrolled for credit flag (stating that the student was enrolled for that semester), academic level (first-year status), primary plan code and the corresponding description (the student's major), for-credit attempted hours for term (the amount of credit hours the student attempted in that semester), online hours attempted for the semester, current semester GPA, gender (dichotomous), first-generation status (if known), race/ethnicity categorization according to IPEDS, age at beginning of semester, high school GPA, number of student life events attended, if known, and if member of a university recognized student club, SAT score, and/or ACT score. There were also a few new variables that were added after doing further investigation of first-year student retention which were related to a student's tuition (discussed in detail later). Data were collected on tuition, mandatory fees, and course/program fees, aid year (for financial aid packaging), housing arrangement (living on or off campus), Pell amount, total grant amount, total loan amount, total scholarship amount, and total financial aid award. These included tuition, fees, grant (including Pell), housing status, and online credit hours.

After the preliminary analysis, additional research and working with the UCCS Department of Institutional Research, the original dataset obtained did not include information around socioeconomic status (SES) which can be identified using financial aid records. Students who are low-income not only have higher barriers to access higher education, but also are retained at a much lower rate (Karimshah et al., 2013). Additionally, as explained in the literature review, housing status (if a student is living on or off campus) is an important indicator of student retention. Both factors were not collected in the original set of variables and were

obtained after submission of the original dataset with IRB and committee approval. Further, employment status was unable to be collected as the system that collects the information was unavailable when all other data was gathered.

After an initial screening of these variables, it was clear there was not a variable specifically indicating whether a student was retained. However, after further investigation it was clear that the answer was embedded in the data. For example, in Table 4.1, a list of pseudo student identification numbers, followed by term code, code description, enrollment flag, and academic level. The dataset was setup so that each student row repeated each time the student enrolled in the subsequent semester. Student 5 began in Fall 2018 and then enrolled again in Spring 2019; whereas students 2, 3, 4, and 6 did not. Student 7 enrolled in Spring 2020, Fall 2020, and again in Spring 2021.

Table 4.1

Raw Enrollment Data

Case Number	Term Code	Term Description	Credit Hours
1	2177	2017 Fall	1
1	2181	2018 Spring	6
1	2187	2018 Fall	6
2	2177	2017 Fall	15
3	2177	2017 Fall	12
4	2187	2018 Fall	9
5	2187	2018 Fall	10
5	2191	2019 Spring	13
6	2187	2018 Fall	5

7	2201	2020 Spring	6
7	2207	2020 Fall	7
7	2211	2021 Spring	6

Therefore, it was assumed that if a student had multiple rows, then they were retained (i.e., student 5 and 7), and if they did not have multiple rows (i.e., cases 2, 3, 4, and 6), then they were not retained. There were a few concerns at this point. For example, it was unknown whether student 5 was retained at the level or was considered to be retained and come back their second fall. It was only known if student 5 enrolled in their second semester (spring term). To understand the full picture of student 5, more data such as the academic standing of sophomores were needed. For many of the students, there were enrollment for fall and spring of their first year but nothing after this because the original request was for freshmen data only. This left the researcher not knowing if that student came back their second year or not. Therefore, the next step was to request the data of all sophomores and then case match them to the current respective dataset. After removing duplicates, the total data set included a total of 9361 individual students. The dependent variable, retention, was created by Mathematics professor Gaetan Delavignette who developed an algorithm using Matlab to match students entering the first semester to their fall semester of their second year.

Next, a complete cases analysis was performed (Table 4.2) to provide a breakdown of the percentage of cases used for each variable. This analysis showed there were specific variables that had a significant amount of data missing. For example, online hours per term had 7823 cases missing (83.6%) of cases. Other notable missing variables included student life events (N = 4753; 50.8%) and student life clubs (N = 2632; 28.1%), financial information such as Pell grants, loans, scholarships, other grants, and total financial aid (N = 1738; 18.6). It is important to note

that missing data could be due to information not being collected or students not needing funding.

Table 4.2

Case Summaries

	Included		Missing		Total	
	N	Percent	*N*	Percent	*N*	Percent
Case number	9361	100.0%	0	0.0%	9361	100.0%
Cohort	9361	100.0%	0	0.0%	9361	100.0%
Credit hours enrolled	9361	100.0%	0	0.0%	9361	100.0%
Online credits hours	1538	16.4%	7823	83.6%	9361	100.0%
Current GPA	9361	100.0%	0	0.0%	9361	100.0%
Age	9361	100.0%	0	0.0%	9361	100.0%
High school GPA	9175	98.0%	186	2.0%	9361	100.0%
Student life events	4608	49.2%	4753	50.8%	9361	100.0%
Tuition and fees	8417	89.9%	944	10.1%	9361	100.0%
Pell award	7623	81.4%	1738	18.6%	9361	100.0%
Grant awards	7623	81.4%	1738	18.6%	9361	100.0%
Loan total	7623	81.4%	1738	18.6%	9361	100.0%
Scholarship total	7623	81.4%	1738	18.6%	9361	100.0%
Financial aid total	7623	81.4%	1738	18.6%	9361	100.0%
SAT score	5010	53.5%	4351	46.5%	9361	100.0%
ACT score	4699	50.2%	4662	49.8%	9361	100.0%
Retention	9361	100.0%	0	0.0%	9361	100.0%

Major	9360	100.0%	1	0.0%	9361	100.0%
First generation status	9289	99.2%	72	0.8%	9361	100.0%
Student life clubs	6729	71.9%	2632	28.1%	9361	100.0%
Housing	7623	81.4%	1738	18.6%	9361	100.0%
Gender	9361	100.0%	0	0.0%	9361	100.0%
Ethnicity	9361	100.0%	0	0.0%	9361	100.0%
Enrollment term	9361	100.0%	0	0.0%	9361	100.0%

Next, a missing variable analysis was conducted to understand which variables were missing and to establish if there were any patterns to the missingness. Figure 4.1 shows out of the 24 variables included in the analysis, 15 (62.5%) had missing data. In the examination of individual cases, a total of 9,325 (99.62 %) cases had at least one missing variable. Out of the total values that were missing (N = 224,664), there were 35,852 (15.96%) that were missing.

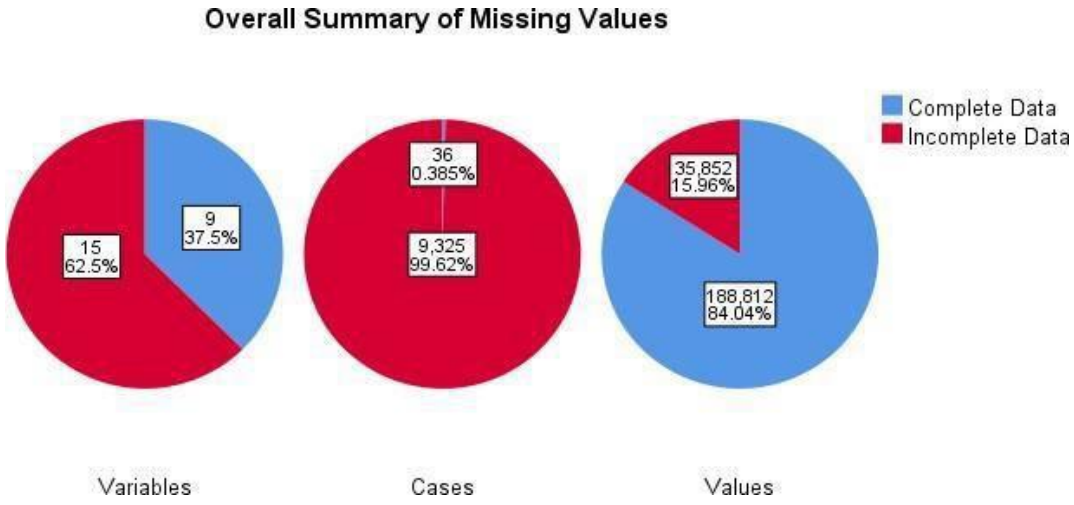


Figure 4.1

Summary of Missing Values

There was a total of 104 patterns of data missingness (Figure 4.2). In a closer examination of the Figure 4.2 represents a specific pattern, and each column represents the

corresponding variable that was missing. For example, moving from left to right, pattern 2 shows that there was one pattern where only first-generation status was missing; whereas, pattern 6 represents a pattern where all financial and housing information was missing. Pattern 104 shows that first generation status, high-school GPA, financial information, housing status, student life events, SAT score, ACT score, and online credit hours were all missing.

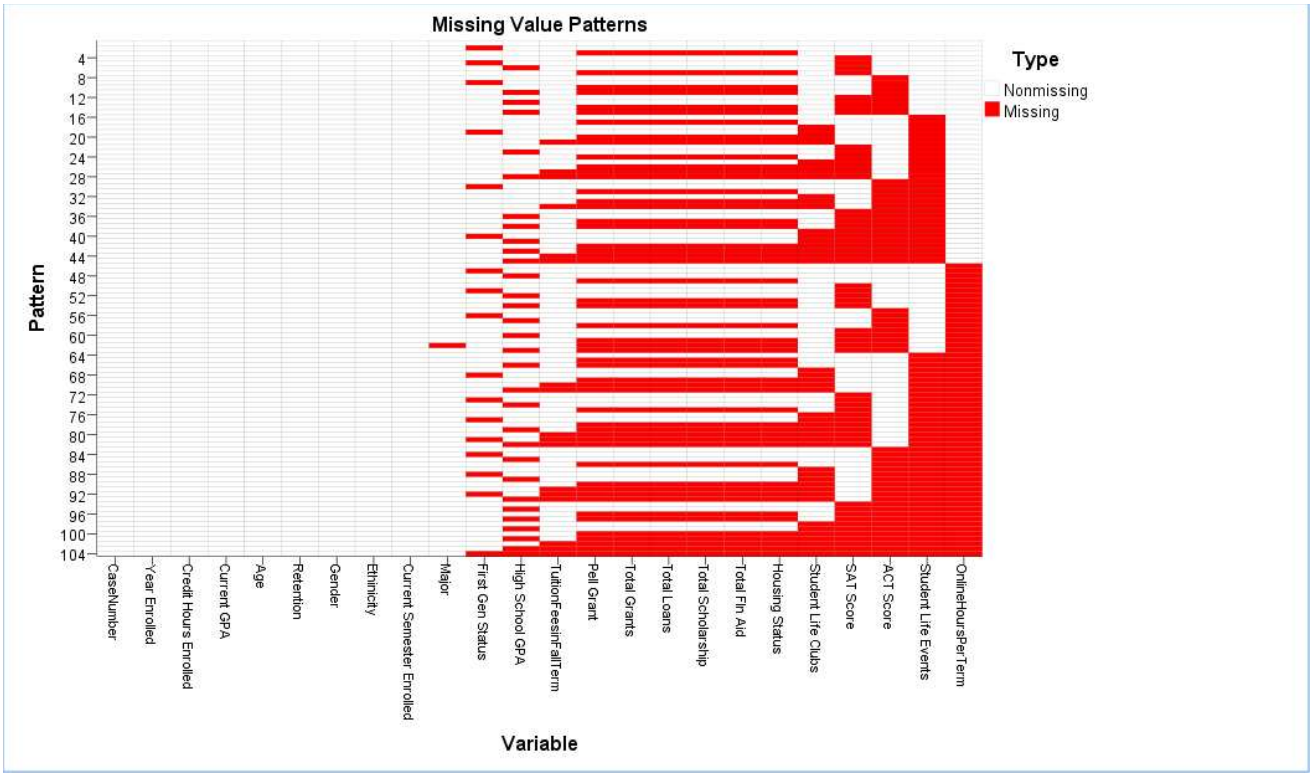
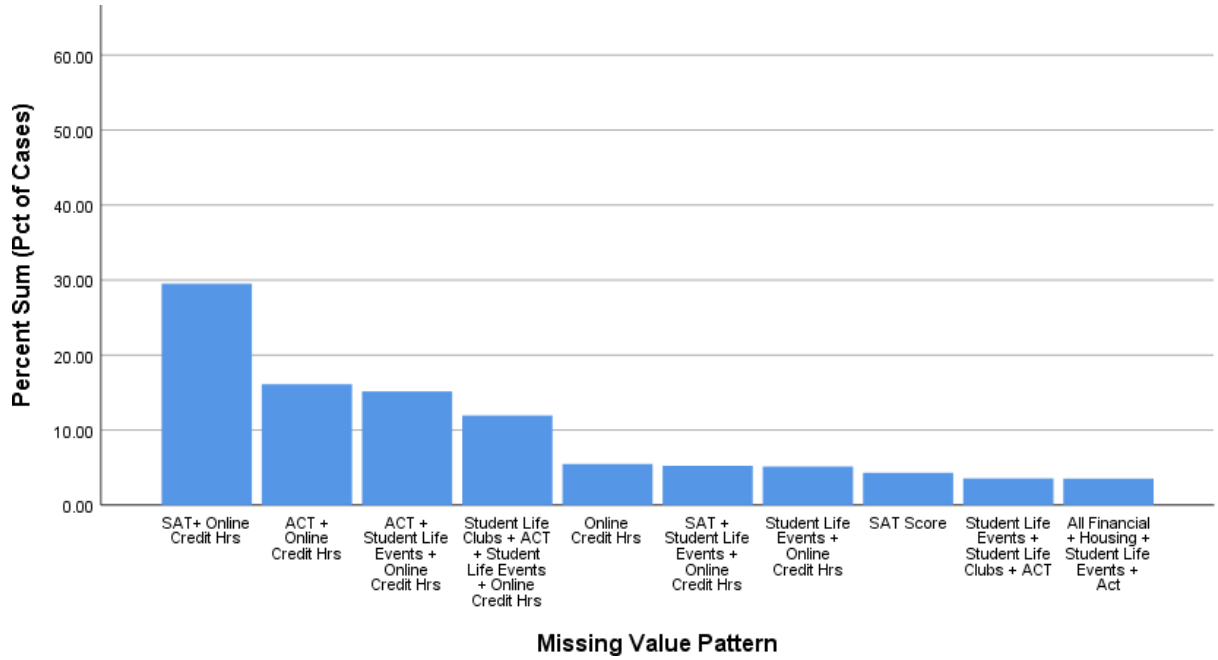


Figure 4.2

Patterns of Missing Values

Next, an analysis was conducted on the percentage of overall missing data by specific missing variable or pattern of variables. Figure 4.3 shows that almost 30% of the missing variables come from the combination of SAT score and online credit hours followed by approximately 17% of ACT score and online credit hours, 15% of the combination of ACT

scores, student life events, and online credit hours, and 12% of student life club participation, ACT score, student life event participation, and online credit hours. Other variables and combinations in the table made up the remaining top 10 most frequently occurring patterns of missing variables. Lastly, the results of Little's (1988) test of missing completely at random was statistically significant, $\chi 2 = 25687.375$, $df = 1703$, $p < .001$; meaning, the data were not Missing Completely at Random (MCAR), and therefore, using single input methods such as listwise deletion should be used with caution.



Note. The 10 most frequently occurring patterns are shown in the chart.

Figure 4.3

Percentage of Missing Values

Listwise Analysis

A logistic regression was conducted to assess whether the aforementioned independent variables significantly predicted student retention. However, when using listwise deletion, there were several complications. Because there were such large amounts of missing data, the model

yielded unusable information when all variables were included in the model. For example, out of the 9,361 total cases in the database, only 0.04% (36 cases) had all of the data for every variable (Table 4.3).

Table 4.3

Summary of Cases (Listwise)

		N	Percent
Selected cases	Included in analysis	36	.4
	Missing cases	9325	99.6
	Total	9361	100.0
Unselected cases		0	.0
Total		9361	100.0

Therefore, the next step for conducting the logistic regression analysis using listwise deletion was to remove the cases that had the most missing data. According to the summary of missing pattern percentages (Figure 4.3), the combination of SAT score and online credit hour accounted for the largest portion of missing data. After removing these two variables, the included cases increased from 0.4% (36 cases) to 28.9% (2674). When all predictor variables were considered together, they significantly predicted whether a student would be retained or not $\chi 2 = 633.06$, $df = 20$, $p < .001$. Table 4.4 shows which specific variables were significant including credit hours, current GPA, student life events and clubs, housing status, major, and first-generation status. The model accounted for between 21% and 29% of the total variance. However, the moderate explanatory power coupled with the low sample size used in this analysis should be viewed with caution in making generations regarding the overall retention outcome.

Table 4.4

Logistic Regression: Excluding SAT Score and Online

	B	*SE*	Odds Ratio	Sig.
Credit hours	0.144	0.025	1.155	0.001
Current GPA	0.896	0.054	2.45	0.001
Student life events	0.114	0.029	1.12	0.001
Tuition and fees	0	0	1	0.001
Major	-0.007	0.002	0.993	0.001
Housing status	0.133	0.042	1.142	0.002
First gen status	0.3	0.099	1.35	0.003
Student club event	0.324	0.124	1.383	0.009
Gender	0.21	0.099	1.234	0.034
Total financial aid	0	0	1	0.068
Cohort	0.132	0.093	1.141	0.156
High school GPA	-0.097	0.099	0.908	0.329
Total loans	0	0	1	0.336
Pell grants	0	0	1	0.397
Age	-0.016	0.029	0.984	0.578
Ethnicity	-0.011	0.021	0.989	0.586
Total scholarship	0	0	1	0.679
Total grants	0	0	1	0.697
ACT score	-0.005	0.015	0.995	0.711
Constant	-269.976	187.754	0	0.15

By removing the next highest missing variable (ACT score) from the model, the included cases increased from 28.9% to 43.4% with no significant changes in model variance or independent predictors. Additionally, after running the logistic regression using listwise substitution, term enrollment was no longer included in the model. Because there were no missing data from this variable, a test of multicollinearity was conducted to investigate any

intercorrelations that could lead to term enrollment not being included. The test concluded that multicollinearity was a concern for cohort and term enrollment (Tolerance = .09, VIF = 10.92). Though these two variables were measuring points of entry to the university (i.e., cohort indicates the first enrollment; term indicates the current enrollment), they were both measuring similar constructs. Therefore, combining both variables into a single construct to develop better internal consistency would be beneficial.

Beta (B) is the expected amount of change in retention (dependent variable) for each one unit change in the predictor. For example, credit hours increase by one unit increase (credit hour), the chances of a student being retained increases by 14%. The odds ratio is calculated in terms of 1 (less than one decreases probability and over 1 increases probability) with the purpose of showing the chances of Beta occurring. Using the same example, the Odds Ratio of credit hours is 1.155 (1.16 rounded) with the odds of students being retained at a 14% increase with every one credit hour increase at 16% (1-1.16). The standard error (*SE*) is the amount of variability that would occur if multiple samples with taken out of the same population. The smaller the number of SE the more precise. Lastly the p-values (Sig). indicates statistical significance (if the value is less than 0.05) and whether to reject the null hypothesis. In this example, credit hours are statistically significant and therefore the null would be rejected to state that credit hours have an impact on student retention.

Using this explanation to understand the rest of the variables in the model, GPA had the highest Beta with the highest odds ratio meaning that if a student is able to increase their GPA they have a substantially higher chance of being retained. Attending an additional student life event increased the chances of being retained at 11% with the odds of this occurring being 12%.

Mean Substitution

The next analysis was mean substitution. Out of the 24 total variables used in the overall analysis, nine had complete data, and therefore, mean substitution was not required. These included case number, cohort, credit hours, current GPA, age, retention, gender, ethnicity, and enrollment term. These variables demonstrated to be vital for non-missing data because using the mean for a categorical variable such as ethnicity was not viable. All missing values were then replaced with the respective mean of that variable (i.e., missing SAT score was replaced with the average of the observed SAT scores). Once this was completed, there were no missing data in the model ($N = 9361$). A logistic regression was then performed on mean substituted data.

When all predictors' variables were considered together, they significantly predicted whether a student would be retained or not $\chi 2 = 3236.87$, $df = 23$, $p < .001$. Table 4.5 shows which specific variables were significant, including all of the variables from listwise deletion as well as financial aid and gender. Further, the model accounted for between 30% and 40% of the total variance.

Table 4.5

Mean Substitution: Variables in Equation

	B	*SE*	Odds Ratio	Sig.
Cohort	.92	.19	2.503	.001
Credit hours	.08	.01	1.085	.001
Current GPA	.79	.03	2.200	.001
Student life events (Mean substitution)	.20	.02	1.216	.001
Housing status (Mean substitution)	.170	.024	1.186	.001
Major (Mean substitution)	-.004	.001	.996	.001

Term enrolled	-.865	.100	.421	.001
Gender	.150	.052	1.161	.004
First generation status (Mean substitution)	.128	.046	1.136	.005
Total financial aid (Mean substitution)	.000	.000	1.000	.008
Ethnicity	-.010	.011	.990	.065
High school GPA (Mean substitution)	.08	.05	1.080	.09
Total loans (Mean substitution)	.00	.00	1.000	.095
Pell grants (Mean substitution)	.00	.00	1.000	.099
Total grants (Mean substitution)	.00	.00	1.000	.235
Total scholarships (Mean substitution)	.000	.000	1.000	.347
Online hours (Mean substitution)	-.02	.02	.985	.52
Student club events (Mean substitution)	.029	.079	1.029	.718
SAT score (Mean substitution)	.000	.000	1.000	.759
Age	-.00	.01	.998	.862
ACT Score (Mean substitution)	-.001	.010	.999	.885
Tuition and fee (Mean substitution)	.00	.00	1.000	.918

However, using mean substitution as the method for handling missing data decreased the level of accuracy the model because classified students were either retained or not retained from 77% to 75%. This was likely due to the large amount of missing data and the need to use the mean for so many values.

Multiple Imputation Analysis

Multiple imputation was the last analysis conducted in the study. This technique examined the patterns in the missing data and replaced them with imputed data created by iterations of the

observed values. This was performed in a two-part series which included an imputation stage and a pooling stage. The imputation stage began with the creation of the baseline iterations for missing data using Mersenne Twister random number generator. Next, the multiple imputation analysis was used to create new dataset of predicted values. The default is five iterations that can be generated to create a model to fit the prediction, and the average of the five is then used as the missing value. The Markov chain Monte Carlo method was then utilized because of the monotonicity and constraints were checked to ensure all data being imputed were plausible (eliminating any outliers). The Markov chain Monte Carlo method provides a specific set of algorithms that are used to sample probabilities distributions. The term 'chain' derives from the notion that each algorithm draws upon the preceding sample for the subsequent analysis (think chain reaction; Brownlee, 2019).

Next, the new dataset was created to encompass iteration history to review iterations of imputed data. The second part of this analysis pooled or aggregated data to run the logistic regression analysis. A total of 9361 cases were used in the model. Using the fifth iteration of variables to predict student retention, the model was statistically significant ($\chi 2 = 3255.09$, $df = 23$, $p < .001$) and predicted between 29% and 39% of the variance in whether a student would be retained using the variables in the model. Table 4.6 specifies the variables that were significant which included many of the same variables as listwise and mean substitution but with slight differences, including the significance of total loans and the marginal significance of gender; whereas previous models showed this was highly significant. Further, using multiple imputation, High school GPA was no longer marginally significant.

Table 4.6

Multiple Imputation: Variables in Equation

	B	*SE*	Odds Ratio	Sig.
Cohort	0.919	0.206	2.507	0.001
Credit hours	0.082	0.012	1.085	0.001
Current GPA	0.792	0.028	2.209	0.001
Student life events	0.130	0.014	1.139	0.001
Major	-0.004	0.001	0.996	0.001
First generation status	0.207	0.049	1.230	0.001
Term enrolled	-0.856	0.107	0.425	0.001
Constant	-1855.870	414.787	0.000	0.001
Housing status	0.165	0.024	1.179	0.001
Total financial aid	0.000	0.000	1.000	0.002
Student club events	-0.177	0.070	0.838	0.012
Total loans	0.000	0.000	1.000	0.039
Gender	0.108	0.064	1.115	0.097
Total scholarships	0.000	0.000	1.000	0.125
Pell grants	0.000	0.000	1.000	0.128
Online credit hours	-0.042	0.027	0.959	0.171
Ethnicity	-0.013	0.011	0.987	0.218
Total grants	0.000	0.000	1.000	0.346
High school GPA	0.047	0.055	1.049	0.390
Age	0.008	0.011	1.008	0.443
Tuition and fees	0.000	0.000	1.000	0.651
ACT score	0.010	0.025	1.010	0.720

SAT score	0.000	0.001	1.000	0.980

Summary of Findings

Table 4.7 is a comprehensive list of the variables that were individually significant within each model, the number of cases that were included, and the total variance of the model. Listwise deletion had the most distinct differences specifically with the cases included. Mean substitution yielded the highest model variance but with the most inaccuracy due to the high percentage of missing variables. For example, although not statistically significant in the model, online credit hours had 83.6% (N = 7823) of the data missing. Because the observed data only consisted of 16.4%, the accuracy of this variable was questionable. Variables that were significant including student life events and financial aid had 20-50% of the data missing. Total loan amount was not significant for listwise deletion or mean substitution but was significant in the multiple imputation model. Described in further detail in Chapter 5, the difference in these models can have significant impact on how higher education administration targets and funds specific student retention programs and populations.

Table 4.7

Summary of Findings

Missing data technique	Significant variables	Cases included	Model variance
Listwise deletion	Credit hours, current GPA, student life events, tuition and fees, major, first-generation status, student club participation, housing status, gender	2674 (28.9%)	21%- 29%
Mean substitution	Cohort, credit hour, current GPA, high school GPA (marginally significant), student life events, total financial aid, major, first-generation status, housing status, gender, term enrollment	9361 (100%)	30%-40%

Multiple imputation	Cohort, credit hour, current GPA, student life events, total loans, total financial aid, major, first-generation status, housing status, gender, term enrollment	9361 (100%)	29% & 39%

CHAPTER FIVE: DISCUSSION

Because there is no universal method to analyze missing data, researchers opt for methods based on their expertise which can cause bias, error, and assumptions about the data. The purpose of this study was to examine the deeper contextual insights into a higher education stakeholder's choice to delete, predict, or substitute missing values in an observed dataset. Empirical research in higher education has examined elements of diversity (demographic differences) in relation to first-year student retention, but few studies have sought to critically understand the quantitative methodologies used to demonstrate these outcomes. The results from this study indicated that listwise deletion produced the most variability, specifically with the cases included. Mean substitution yielded the highest model variance but with the most inaccuracy due to the high percentage of missing variables; whereas multiple imputation produced the most accurate results but used an algorithm unlikely to be understood by practitioners in the field of higher education.

This chapter contains discussion and future research possibilities to help answer the following research overarching questions:

1. **(R1):** Does using different methodological techniques for missing values in each dataset impact the interpretation of statistical outcome?
2. **(R2):** How does the overall prediction model change depending on how missing data are inputted?

This chapter includes a discussion of major findings within each model and the practical connections to real world applications in higher education retention interventions. Also included is a discussion on connections to this study and theories such as quantitative critical race theory

and critical data studies. The chapter concludes with a discussion on limitations as well as recommendations for future practice

Interpretation of the Findings

Each model used the same data and predictor variables to explain first-year student retention. The method used to fill the void of any data that was missing was the only difference in each model. The first model removed any student that had missing data, the second used the average of the observed variable in the specific category to replace any missing values, and the third model multiplied the dataset many times to substitute missing values with a predicted data point generated from the iterations of imputed variables. All three models predicted over 25% of the variance but yielded key differences described in detail in this chapter.

Listwise Deletion and the Counternarrative

Listwise deletion asserts that any observation with at least one missing value be excluded from the overall analysis. This technique is often the default method in most statistical packages but has been widely used in quantitative methodologies for its simplicity. Research stated that listwise deletion can be viewed as testing the statistical power of data, with high percentages of missing data causing concern for decision errors (type I and type II; Pepinsky, 2018). Substantial amounts of missing observations mean decreased accuracy to draw conclusions about a population using sample data (Cohen, 1992).

In this study, over 99% of the students had more than one missing value which reduced the dataset from 9,361 students in the model to 36 students (1%). Table A.1 in Appendix A represents the breakdown of the 36 students included in the initial model. Although this data is not sufficient to make generalizable claims due to the small sample size, it is important to see

who is represented in this model. Future research may be more persistent in their pursuit of inclusive variables that capture a diverse sample in both content and context.

The GPA, age, financial aid information, and SAT/ACT score ranged substantially from those students who were retained from those who were not retained. However, six out of the 11 students who identified as a person of color were not retained and only three females of color were retained. All but one student was taking a full course load (over 12 credit hours in per semester). Grade point average (GPA) was lowest for the student (case 924) who was taking less than 12 credits and they were at the top of the age range at 21. Student 924 was a White male who had a 4.0 GPA in high school, did not receive any financial aid (which suggests they had a high socioeconomic background), had a high ACT and SAT score, but did not attend any events on campus. From viewing these data alone, it is plausible that because the student did not attend any student life events or join any clubs, they felt that they did not belong and therefore began to do poorly in class which resulted in them not being retained. While this assumption may seem like a leap, there is research supporting active engagement and its relation to student retention (Astin, 1999). However, many of the other students who were not retained had attended at least one student life event or was part of a club on campus. Other findings showed that all students in the model took some online credits; although over 83% had this variable missing in the larger dataset.

The small percentage of data analyzed in the original listwise deletion removed this method as a viable option for the overall analysis; however, if there were more data available, listwise deletion would be treated as a feasible method of handling missing values. Methodologists often refer to the common practice of having at least 10 cases for each

independent variable (Kaliyadan & Kulkarni, 2019). In the current study this would mean a sample of 230 because there are a total of 23 independent variables.

The next stage in the listwise deletion analysis removed the two variables with the highest amount of missing data, which was online credit hours and SAT scores, with the purpose of increasing the sample size. After removing these two variables, the observations increased from 0.4% (36 cases) to 28.9% (2674). Statistically, this supported listwise analysis, and it also eliminated a potentially critical narrative that was then removed in understanding online credit hours and standardized tests scores in the model. This also showed potential bias in researcher analysis through the removal of variables that were hindering the most generalizable result. Between the novel literature on distance learning coupled with the move to change all higher education courses to remote learning during the past year due to the global pandemic, removing a vital variable such as online credits could impact the retention interventions such as exploring digital literacy and access (Ali, 2020).

The removal of variables due to large amounts of missing data removes vital information that could have serious implications on missed opportunities for meaningful targeted interventions. Even though this may seem like an outdated form of handling missing data, it is still being used as a form of data cleansing. In a previously mentioned study by Delen (2010), the researcher excluded anomalies and any variables they did not find useful. For example, Delen "removed all international student records from the dataset because they did not contain some of the presumed important predictors (e.g., high school GPA, SAT scores)" (p. 501).

Although this explicit form of exclusion occurs less often, it is still a common practice in research. Higher education literature uses language such as, "241 other students enrolled in this course who were not eligible to participate in this study" (Canning et al., 2018, p. 837); whereas

other research breaks down the removal of data such as the study by Han et al. (2017) which examined the impact of mindset on retention. Han et al. stated, “after exclusions, the final sample size was 1,400 students, reflecting 45% of the entering Fall 2013 class (3,104 students)” (p. 1125). The reasons for exclusion were unsigned consent forms, academic status, or lack of permission from course instructors. In the words of Harel et al. (2008), listwise deletion is “a method that is known to be one of the worst available” (p. 351) even though it is often considered the de-facto technique to cleanse data of missing information (Myers, 2011). On the one hand, it is reasonable to remove data that are either not relevant to the study or contains substantial missing information; on the other hand, it is important for a researcher to ask themselves why that data are left out or missing in the first place. For example, unsigned consent forms could raise concerns about a student’s home life and lack of instructor consent could later prove bias in their grading methods (Sablan, 2019).

Preservation, Erasure, and Mean Substitution

Out of the 24 variables included in the model, 15 had missing information that was substituted for the mean. Table 5.1 shows which variable had missing data: online hours, high school GPA, student life events, tuition and fees, Pell grants, total grants, total loans, total scholarships, total financial aid, SAT score, ACT Score, major, first-generation status, student club events, and housing status. When mean substitution occurs for observed values that do not have a lot of missing values (i.e., high school GPA and major), a sample is still preserved which means it can still be representative of the overall population. However, when there is substantial data that are missing (i.e., online credit hours, student life events, SAT/ACT scores), there is a higher risk of the data being biased because there is more information being replaced than what is currently being represented.

Table 5.1

Variables with Missing Values

	Included		Missing		Total	
	N	Percent	N	Percent	*N*	Percent
Online credits hours	1538	16.4%	7823	83.6%	9361	100.0%
High school GPA	9175	98.0%	186	2.0%	9361	100.0%
Student life events	4608	49.2%	4753	50.8%	9361	100.0%
Tuition and fees	8417	89.9%	944	10.1%	9361	100.0%
Pell award	7623	81.4%	1738	18.6%	9361	100.0%
Grant awards	7623	81.4%	1738	18.6%	9361	100.0%
Loan total	7623	81.4%	1738	18.6%	9361	100.0%
Scholarship total	7623	81.4%	1738	18.6%	9361	100.0%
Financial aid total	7623	81.4%	1738	18.6%	9361	100.0%
SAT score	5010	53.5%	4351	46.5%	9361	100.0%
ACT score	4699	50.2%	4662	49.8%	9361	100.0%
Major	9360	100.0%	1	0.0%	9361	100.0%
First generation status	9289	99.2%	72	0.8%	9361	100.0%
Student life clubs	6729	71.9%	2632	28.1%	9361	100.0%
Housing	7623	81.4%	1738	18.6%	9361	100.0%

For example, in the current study, nearly 8,000 students had their online credit hours information artificially created by the observed dataset (less than 1,600). This not only decreased individual variability and but also could change variable characteristics (i.e., variance, median; Little & Rubin, 1989); furthermore, it could result in standard errors that are too low, which increases the chances of Type I (false positive) errors (Béland et al., 2018). Preserving the data are usually in the best interest of the researcher; however, it can erase any anomalies, outliers, or underrepresented demographics which can subsequently perpetuate oppressive regimes of historically excluded populations (Zuberi, 2001). For example, Covarrubias (2011) showed how aggregated state census data concealed the intersectional impact of educational outcomes based on "gender-based discrimination, patriarchy, class inequality, nativist racism" (p. 103) within the Latinx community. Similarly, research from Hogan (2017) used the American Community Survey (ACS) to demonstrate how current reporting structures inflated the grouping of people who identified as having Hispanic origin and race. This led to the omission of data that showed "poverty rates among Latinas/os/x identifying as white [*sic*] which are consistently lower than among Latinas/os/x identifying as 'some other race' or Black/AfroLatinas/os/x2" (Garcia et al., 2018, p. 153). Therefore, when critical variables are missing, the entire narrative is reframed to a master narrative that forces on dominant ideologies and less from marginalized populations.

Importance of Sample Distribution

In the current study, over 50% (4753) of the data on student life events were missing. Of the observed data, a total of 2,279 students attended at least one event on campus with less students attending at higher event rates. Figure 5.1 shows the minimum event participation of zero (n = 2527) and maximum of 22 (n = 1). Importantly, the difference between zero (no one attending) and missing (unknown amount) is about what information is being extracted for

reporting and for what purpose (to be discussed further in below in recommendations). A student can attend zero events throughout the semester if this is still reported and not treated as missing. Instances such as this, as well as other information with true zeros (i.e., financial records), need to be considered when reporting results. Unless the research is able to report these differences, they may be grouped together which can alter the outcomes.

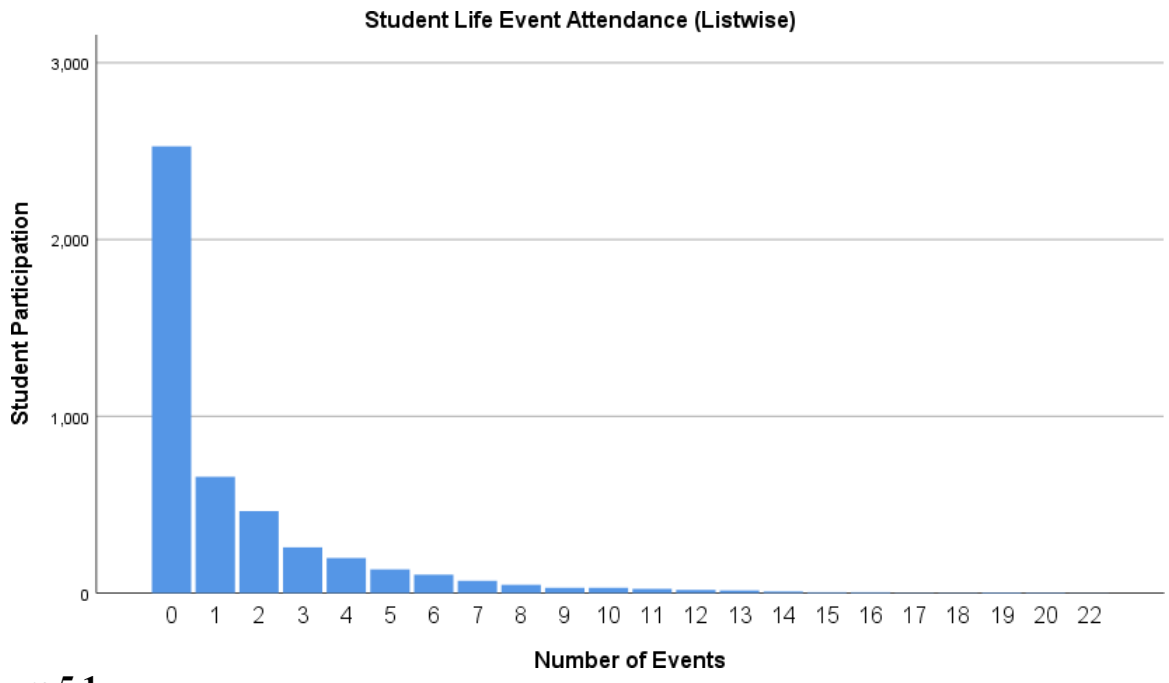


Figure 5.1

Student Life Attendance: Listwise Deletion

Using mean substitution to examine student life events (Figure 5.2), the minimum and maximum range were the same as the listwise deletion results. However, in closer examination, the standard error changed substantially (Listwise *SE* = 0.40; mean substitution *SE* = 0.02). This difference indicated there was more variability by completely removing any missing variables and using the mean to replace any omitted values. The results indicate that a majority of students attended some event on campus (*N* = 4,753; 84.8%). However, the accuracy of this information was questionable because it may not have been truly representative of those that were missing

data. Listwise may have had more variability, but it was more representative of the overall population, and mean substitution provided a higher sample population that was less inclusive of those students who had omitted data.

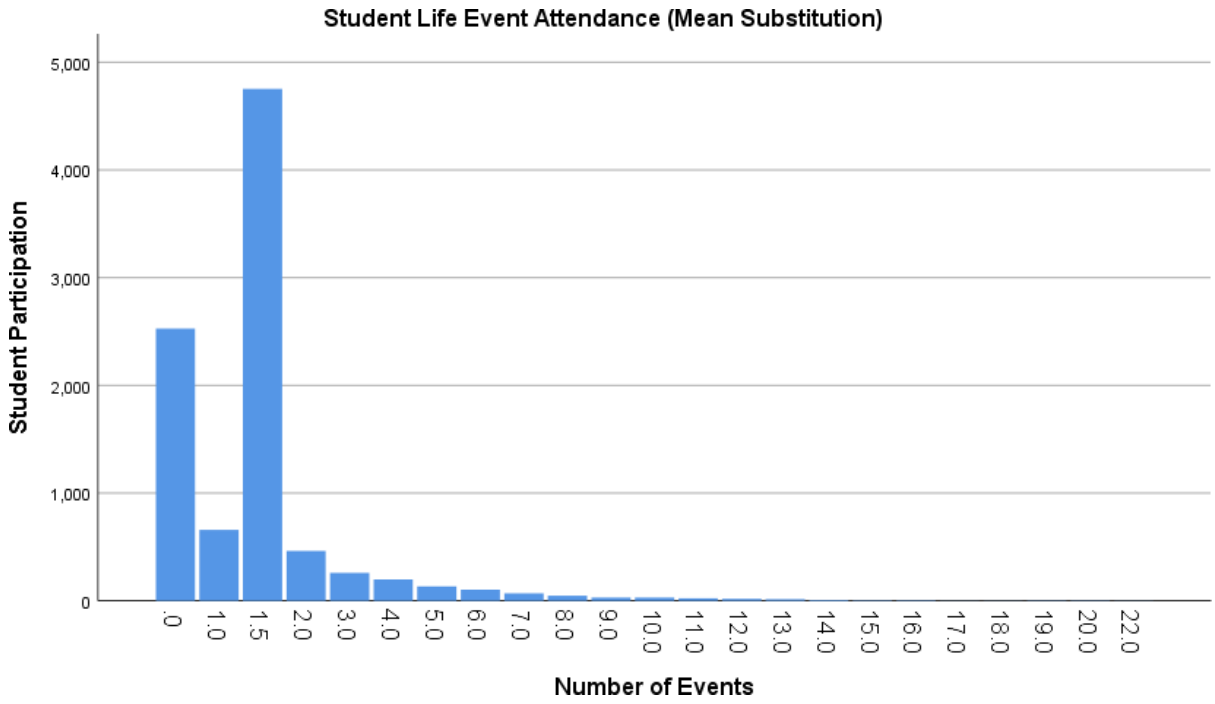


Figure 5.2

Student Life Event Attendance

The differences in these two sets of results painted a very different picture of students who participated on campus. Although using very different types of information, the current study example reflected the work of Hogan (2017) and Covarrubias (2011) by showing how aggregating information can whitewash important data that can be used to share critical information for specific student populations. For instance, if a majority of the students who had missing data for student life event participation did not actually attend any events, but mean substitution concluded they did and they were still not retained, there may not be a push to engage students on campus as an early retention intervention. This could impact program

development, funding, and resource allocation for this division of student affairs because administration may see campus engagement as an effective way to retain students.

Multiple Imputation: Black Box Interpretations

The last method used in the analysis was multiple imputation. The aim of this approach was to predict missing values through a series of imputed data sets that resulted in a combination of multiple iterations as the final outcome (Rubin, 1977). The benefit of this method was the repetitive imputation of predicted values to obtain a more accurate standard error (e.g., how close your sample is the to general population). In contrast, mean substitution introduces biases by using only the observed data as the substitute for the missing values or listwise deletion, which removes all cases with missing data (and substantially decreases the sample size depending on the amount of missing information). Multiple imputation keeps the sample size intact while also maintaining generalizability (Reiter et al., 2006). This method is the most often used among statisticians and data scientists but not commonly used in the field of education (Cox et al., 2004). One reason is due to the practical limitations, such as using this method on large datasets which could take hours or days to complete. Second, this method increases the likelihood of random error (i.e., inaccurate measurements caused by human miscalculations; Goldstein, 2018).

Lastly, the computational complexity required to employ this method and interpret the results in a meaningful way is often lost in translation between the data scientist and the practitioner. This analysis contributes to the opaqueness of the black box problem by providing predictive mean matching using a complex algorithm that lacks the ability to be explained without having prior knowledge. The lack of methodological transparency, specifically when it comes to analyzing student retention information that employs critical implications, begin to lose meaning and value when they are not fully understood by its users (Boyd & Crawford, 2012).

Study Limitations

Several limitations should be noted in this research. First, the current study was conducted using data from a predominantly White institution where enrollment numbers generally tend to be skewed toward individuals who identify as Caucasian. The data used were reflective of this with over 60% ($N = 5599$) of the sample population consisting of White students. The sample may create a bias by developing a master narrative using the majority demographic as the baseline for data that could potentially be used to implement policies that consequently impact students from historically excluded backgrounds (Premraj et al., 2019).

Next, there was a substantial number of variables that had missing information due to the lack of a centralized reporting structure. The university in the current study utilized many complex data sources to accumulate substantial amounts of information that were used to analyze trends and disseminate reports to stakeholders at state and federal levels. With data sources ranging from web-based tools for accessing reports, self-reported data from enrollment applications (i.e., first generation status), to student card swipes at events to measure engagement, incomplete and missing information became more prevalent and difficult to control.

Lastly, the study was bound by the techniques and data used in the analysis. The current study only used three specific missing data techniques and compared their results. Although these methods were selected because they not only vary greatly from each other, but they also are widely used in the field of data analytics, more research should be conducted on the use of maximum likelihood, pairwise deletion, and other commonly used techniques for working with missing data (Chetverikov, 2019).

Recommendations for Future Practice

This research study contends that truth and interpretation are not mutually exclusive, and therefore, must be carefully considered in order to emancipate the societal structures of oppressive regimes that are embedded in everyday culture. The only way to uncover potential inequities is to critically interrogate the systems that have been historically used to create them. Therefore, this study offers three recommendations in the following sections.

Recommendation 1: Contextualizing Data Missingness

Higher Education literature often explains the final sample size in the research, but it does not delve into why these data may have been missing and the impact they may have on the final results (O'Neil, 2016). Peugh and Enders (2004) examined the leading educational journals from 1999 to 2003 and found that out of the 389 reviewed, all but six studies either completely ignored data that were missing or addressed them minimally. A more recent study found that of 20 articles published in 2012 by the *Review of Higher Education*, all but one appeared to have no missing data and yet only three provided explicit justification for the missingness and how it was addressed (Cox et al., 2014). Although missing data are inevitable in quantitative educational research, the decision to "often ignore – a problem for which there is no perfect solution" (Cox et al., 2014, p. 4) is not ideal nor recommended. Concerns regarding data missingness (i.e., what is missing and why) need to be addressed head on to avoid erasure of the counter stories that occur when information is not collected or accounted for. The narrative behind the data is needed to develop equity because numbers do not speak for themselves, and it is the responsibility of the researcher to unpack the dimension of injustice by examining who is being included and why (Stone, 2013).

Recommendation 2: Centralized Reporting Structures

Burdensome processes have made finding and resolving incomplete information more difficult to control or anticipate. When examining factors that impact student retention, data scientists often piecemeal information together from various web-based and human-controlled systems. This information is then gathered and inputted into a large database that is used in the final reporting to campus, state, and federal systems. A critical concern is that each of these stakeholders may require different ad-hoc data requests that rely on specific departments or systems that may or may not be connected. For example, information regarding student financial records is not in the same system as the registrar (i.e., credit load and course platform) or student life events. The responsibility then falls on the data scientists to aggregate the information into a coherent narrative to report out. For example, students who fit a specific profile regarding engagement, credit load, and financial status are more or less likely to be retained. Therefore, it is recommended to have a system in place that each department inputs data into that can pull a master report for the data scientists to review. This could alleviate the burden of merging datasets and potentially eliminate bias that could be introduced when humans input data. However, it should be noted that although data would be inputted into the central system by humans (and therefore subject to error), data have a greater chance of being authenticated by all parties who possess access which could alleviate potential bias.

Recommendation 3: Acknowledging Computational Reflexivity

Recognizing researcher-centered positionality is becoming more common for both qualitative and quantitative research (Secules et al., 2021). The way a researcher views the world can impact how their research is conducted and the conclusions that are established, but why specific methods were used over another is not discussed. A logistic regression may be used over

a multiple regression because the dependent variable is dichotomous and not categorical; however, there is still a significant amount of flexibility and freedom to choose certain methods over others. Specifically, it is recommended that researchers begin to share their worldview, approach, or positionality to explain the reasoning behind including or omitting certain cases, variables, or values. Unpacking the use of specific methods can be viewed as the development of a "cultural toolkit" (Reyes, 2020, p. 221) of statistical techniques that situate the context of reflexivity, or the ways in which an individual's social position can impact their judgments during the research process (Reinhart & Reuland, 1993). This provides the reader more insight on (a) why the researcher chose to study what they did and (b) why they chose the methods that resulted in specific outcomes.

Conclusion

Why does missing data matter? Data are another indication of power: The power to remain hidden. Suppressing specific information allows the dominant culture to manifest itself in the statistics that are supposed to be used to dismantle systems of oppression for underrepresented populations and communities. Personal values and beliefs are embedded into data collection, variable selection, information omission, output interpretation, and overall transparency. Just as datafication can disguise inequities and systems of oppression, so can its absence. Flyverbom et al. (2016) introduced the term "visibility management" (p. 98) which is the level of salience to describe how digital technologies are seen, known, and regulated and the interplay between knowledge and power. The ability to make data visible (i.e., specific variable collection or data omission) involves a form of control around the transparency, disclosure, and accountability of information that are deeply dependent on the acting individual or organization.

The current study situates the findings of the research within the five tenets of quantitative critical race theory (Gillborn et al., 2018) by stating that (1) Centrality in racism is created by employing the oppressive systems that gather information for subsequent data analysis. This is seen in using data from a predominately White institution and the erasure of counter narratives through data omission (2) Numbers are not neutral and the researcher's choice to use specific methods for handling missing data without explanation lacks transparency (3) Categories are socially constructed which is evident through the discrepancies in federal definitions of variables such as retention (4) Voice and insight where those who interpret and disseminate information provide the knowledge that is used to create or hinder change, and lastly (5) Creating social justice with numbers by diving into methodologies that are often overlooked (such as examining data missingness) to move towards more equitable practices.

By examining multiple databases using various methods for missing data, the current study should be replicated to provide further information that fitting algorithms with data do not change the data; it changes the interpretation of the result which cannot be "divorced from the social contexts in which these technologies are situated" (Elish & Boyd, 2017, p. 19). This research examined the opaqueness of the black box data methods by understanding the practical applications and deeper contextual insights of statistical methodologies and prove how minor methodological changes, such as how one handles missing data, can have larger implications on not only the statistical outcome but the impact on actionable items within systems like policy, program developments, and targeted outreach for proactive interventions (Boyd & Crawford, 2012; Yousif, 2015).

REFERENCES

Ali, I. (2020). The covid-19 pandemic: Making sense of rumor and fear: Op-ed. *Medical Anthropology*, *39*(5), 376-379. https://doi.org/10.1080/01459740.2020.1745481

Aliyeva, A., Cody, C. A., & Low, K. (2018). *The history and origins of survey items for the integrated postsecondary education data system (2016–17 Update)*. U.S. Department of Education, National Postsecondary Education Cooperative. http://nces.ed.gov/pubsear

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies*, *6*(2), 1-18. https://doi.org/10.5539/hes.v6n2p1

Allison, P. D. (2002). *Missing data* (Vol. 136). Sage Publications.

Altunbaş, Y., & Thornton, J. (2020). Finance and income inequality revisited. *Finance Research Letters*, *37*, 101355. https://doi.org/10.1016/j.frl.2019.101355

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/

Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response rates in organizational science, 1995–2008: A meta-analytic review and guidelines for survey researchers. *Journal of Business and Psychology, 25*(3), 335–349. https://doi.org/10.1007/s10869-010-9157-6

Asghar, J. (2013). Critical paradigm: A preamble for novice researchers. *Life Science Journal, 10*(4), 3121-3127.

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel*, *25*(4), 297–308.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing dataanalyses. *Journal of School Psychology*, *48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Barbera, S. A., Berkshire, S. D., Boronat, C. B., & Kennedy, M. H. (2020). Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. *Journal of College Student Retention: Research, Theory & Practice*, 22(2), 227-250. https://doi.org/10.1177/1521025117738233

Béland, J. P., Bernier, L., Dagenais, P., Daniel, C. É., Gagnon, H., Parent, M., & Patenaude, J. (2018). Revisiting the fact/value dichotomy: A speech act approach to improve the integration of ethics in health technology assessment. *Open Journal of Philosophy, 8,* 578-593. https://doi.org/10.4236/ojpp.2018.85042

Berry, D. (2011). The computational turn: thinking about the digital humanities. *Culture Machine, 12*. https://culturemachine.net/wp-content/uploads/2019/01/10-Computational-Turn-440-893-1-PB.pdf

Bichsel, J. (2012). Analytics in higher education benefits, barriers, progress, and recommendations (Research report). *Research Gate*. https://doi.org/10.13140/RG.2.1.1064.6244

Bland, M. (2015). *An introduction to medical statistics*. Oxford University Press.

Blue, A. (2018). *Researcher looks at 'digital traces' to help students*. AU News. https://uanews.arizona.edu/story/researcher-looks-digital-traces-help-students

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5),662-679. https://doi.org/10.1080/1369118x.2012.678878

Braxton, J. M., Sullivan, A., S., & Johnson, R. T. (1997). Appraising Tinto's theory of college student departure. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (Vol. 12, pp. 107-158). Agathon.

Brownlee, J. (2019, November 19). A gentle introduction to Markov Chain Monte Carlo for probability. *Probability*. https://machinelearningmastery.com/markov-chain-monte-carlo-for-probability/

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly, 101*(2). https://doi.org/10.1111/papq.12302

Bruckner, M. A. (2018). The promise and perils of algorithmic lenders' use of big data. *Chi.-Kent L. Modern Law Review.,* 93, 3. https://doi.org/10.1111/1468-2230.12316

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Canning, E. A., Harackiewicz, J. M., Priniski, S. J., Hecht, C. A., Tibbetts, Y., & Hyde, J. S. (2018). Improving performance and retention in introductory biology with a utility-value intervention. *Journal of Educational Psychology*, *110*(6), 834. https://doi.org/10.1037/edu0000244

Capers, Q., Clinchot, D., McDougle, L., & Greenwald, A. G. (2017). Implicit racial bias in medical school admissions. *Academic Medicine*, *92*(3), 365-369. https://doi.org/10.1097/ACM.0000000000001388

Carolan, M. (2017). Publicising food: big data, precision agriculture, and co-experimental techniques of addition. *Sociologia Ruralis, 57*(2), 135-154. https://doi.org/10.1111/soru.12120

Carnevale, A. P., Smith, N., & Strohl, J. (2013). Recovery: Job growth and education requirements through 2020. Georgetown University Center on Education and the Workforce. https://cew.georgetown.edu/cew-reports/recovery-job-growth-and-education-requirements-through-2020/

Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). *America's divided recovery: College haves and have-nots. Georgetown University Center on Education and the Workforce*. https://cew.georgetown.edu/wp-content/uploads/Americas-Divided-Recovery-web.pdf

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access, 5*, 8869-8879. https://doi.org/10.1109/access.2017.2694446

Chetverikov, D. (2019). Testing regression monotonicity in econometric models. *Econometric Theory, 35*(4), 729-776. https://doi.org/10.1017/s0266466618000282

Chhabra, G., Vashisht, V., & Ranjan, J. (2017). A comparison of multiple imputation methods fordata with missing values. *Indian Journal of Science and Technology, 10*(19), 1-7. https://doi.org/10.17485/ijst/2017/v10i19/110646

Conaway, W., & Bethune, S. (2015). Implicit bias and first name stereotypes: What are the implications for online instruction? Online Learning Journal, *19*(3), 162–178. http://dx.doi.org/10.24059/olj.v19i3.674

Corlett, S., & Mavin, S. (2018). Reflexivity and researcher positionality. In C. Cassell, A. L. Cunliffe, & G. Grandy (Eds.), *The SAGE handbook of qualitative business and management research methods: History and traditions* (pp. 377-399). Sage Publications. https://dx.doi.org/10.4135/9781526430212.n23

Covarrubias, A. (2011). Quantitative intersectionality: A critical race analysis of the Chicana/o educational pipeline. *Journal of Latinos and Education, 10*(2), 86-105.

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer and real-world example. *The Review of Higher Education*, *37*(3), 377-402. https://doi.org/10.1353/rhe.2014.0026

Crosling, G. (2017). Student retention in higher education, a shared issue. In P. Nuno Teixeira, J-C. Shin, A. Amaral, A. Bernasconi, A. M. Magalhaes, B. M. Kehm, B. Stensaker, E. Choi, E. Balbachevsky, F. Hunter, G. Goastellec, G. Mohamedbhai, H. de Wit, J.

Välimaa, L. Rumbley, L. Unangst, M. Klemencic, P. Langa, R. Yang, & T. Nokkala, (Eds.), *Encyclopedia of international higher education systems and institutions* (pp. 1-6). Springer. https://doi.org/10.1007/978-94-017-9553-1_314-1

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, *46*(5), 904-920. https://doi.org/10.1007/978-3-319-06520-5_1

Dalton, C., & Thatcher, J. (2014, May 12). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space, 29*. https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Dean, J., Furness, P., Verrier, D., Lennon, H., Bennett, C., & Spencer, S. (2018). Desert island data: an investigation into researcher positionality. *Qualitative Research*, *18*(3), 273-289. https://doi.org/10.1177/1468794117714612

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195-210. https://doi.org/10.3386/w8432

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506. https://doi.org/10.1016/j.dss.2010.06.003

Delgado, R., & Stefancic, J. (2017). *Critical race theory*. New York University Press.

DeWitz, S. J., Woolsey, M. L., & Walsh, W. B. (2009). College student retention: An exploration of the relationship between self-efficacy beliefs and purpose in life among college students. *Journal of college student development*, *50*(1), 19-34. https://doi.org/10.1353/csd.0.0049

Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. (2013). The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes. National Bureau of Economic Research Working Paper Series. https://doi.org/10.3386/w19552

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: A systematic review of how they are reported and handled. *Epidemiology*, *23*(5), 729-732. https://doi.org/10.1097/ede.0b013e3182576cdb

Elish, M. C., & Boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication monographs, 85*(1), 57-80. https://doi.org/10.1080/03637751.2017.1375130

Enders, C. K., & Baraldi, A. N. (2018). Missing data handling methods. In P. Irwing, T. Booth,

D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 139-185). Wiley & Sons. https://doi.org/10.1002/9781118489772

Erwin, P. G. (2006). Children's evaluative stereotypes of masculine, feminine, and androgynous first names. *The Psychological Record, 56*, 513–519. https://doi.org/10.1007/bf03396031

Falk, C. F., & Heine, S. J. (2015). What is implicit self-esteem, and does it vary across cultures? *Personality and Social Psychology Review, 19*(2), 177–198. https://doi.org/10.1177/1088868314544693

Fazio, R. H., Jackson, J. R., Dutton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27*(4), 307-316. https://doi.org/10.1207/s15324834basp2704_3

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*(1), 74-147. https://doi.org/10.1080/10463280600681248

Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community college review, 36*(2), 68-88. https://doi.org/10.1177/0091552108320222

Flyverbom, M., Leonardi, P., Stohl, C., & Stohl, M. (2016). Digital age: The management of visibilities in the digital age. *International Journal of Communication,* 10, https://doi.org/12. 1932–8036/20160005

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management, 36*(3), 403-413. https://doi.org/10.1016/j.ijinfomgt.2016.01.006

Fuller, C. (2011). *The history and origins of survey items for the integrated postsecondary education data system*. U.S. Department of Education, National Postsecondary Education Cooperative. http://nces.ed.gov/pubsearch

Garcia, J. (2019, October 25). Cultivating student success: Maintaining equity in education. National Center for Developmental Education Conference. Colorado Springs, Colorado.

Garcia, N. M., López, N., & Vélez, V. N. (2018). QuantCrit: Rectifying quantitative methods through critical race theory. *Race Ethnicity and Education*, *21*(2), 149–157. https://doi.org/10.1080/13613324.2017.1377675

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry, 28*(4), 268–272. https://doi.org/10.1080/1047840x.2017.1373546

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451-482. https://doi.org/10.1146/annurev-psych-120709-145346

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: education, policy, 'big data' and principles for a critical race theory of statistics. *Race Ethnicity and Education, 21*(2), 158-179. https://doi.org/10.1080/13613324.2017.1377417

Gitelman, L. (2013). *Raw data is an oxymoron*. MIT press.

Golbeck, J. (2016). User privacy concerns with common data used in recommender systems. In E. Spiro & Y. Y. Ahn (Eds.), *SocInfo 2016: Social informatics: Lecture notes in computer science* (Vol. 10046, pp. 468-480). Springer. https://doi.org/10.1007/978-3-319-47880-7_29

Goldstein, M. (2018). Has New York been promoting unreliable DNA evidence leading to wrongful convictions? *Syracuse Journal of Science & Technology, 35(3).*

Gorelick, M. H. (2006). Bias arising from missing data in predictive models. *Journal of Clinical Epidemiology*, *59*(10), 1115-1123. https://doi.org/10.1016/j.jclinepi.2004.11.029

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197-218. https://doi.org/10.1207/s15327906mbr3102_3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464 –1480. https://doi.org/10.1037/0022-3514.74.6.1464

Guba, E. G. (Ed.). (1991). The alternative paradigm dialogue. In *The paradigm dialogue* (pp. 17-30). Sage.

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018, January 18). *The promise of performance assessments: Innovations in high school learning and college admission*. Learning Policy Institute. https://learningpolicyinstitute.org/product/promise-performance-assessments-report

Hagood, L. P. (2019). The financial benefits and burdens of performance funding in higher education. *Educational Evaluation and Policy Analysis, 41*(2), 189-213. https://doi.org/10.3102/0162373719837318

Han, C.-W., Farruggia, S. P., & Moss, T. P. (2017). Effects of academic mindsets on college students' achievement and retention. *Journal of College Student Development, 58*(8), 1119-1134. https://doi.org/10.1353/csd.2017.0089

Harrell, F. E. (2015). Missing data. In *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (pp. 45-61). Springer.

Harel, O., Zimmerman, R., & Dekhtyar, O. (2008). Approaches to the handling of missing data in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The SAGE sourcebook of advanced data analysis methods for communication research* (pp. 349-371). Sage Publications. https://dx.doi.org/10.4135/9781452272054.n12

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology* (pp. 1-20). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119125563.evpsych241

Heshusius, L. (1994). Freeing ourselves from objectivity: Managing subjectivity or turning toward a participatory mode of consciousness? *Educational Researcher*, *23*(3), 15-22.

Hill, K. (2012). How Target figured out a teen girl was pregnant before her father did. *Forbes.* https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=7023103e6668

Hogan, H. (2017). Reporting of race among Hispanics: Analysis of ACS data. In *The frontiers of applied demography* (pp. 169-191). Springer. https://doi.org/10.1007/978-3-319-43329-5_9

Holmes, A. G. D. (2020). Researcher positionality--A consideration of its influence and place in qualitative research--A new researcher guide. *Shanlax International Journal of Education, 8*(4), 1-10. https://doi.org/10.34293/education.v8i4.3232

Hopp, S. L., Owren, M. J., & Evans, C. S. (Eds.). (2012). *Animal acoustic communication: Sound analysis and research methods*. Springer Science & Business Media.

Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis, 54*(12), 3095-3107. https://doi.org/10.1016/j.csda.2009.11.023

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society, 3*(2), https://doi.org/2053951716674238

Integrated Postsecondary Education Data System. (2020). Undergraduate retention and graduation rates. https://nces.ed.gov/programs/coe/indicator_ctr.asp

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Joselson, N. (2016, October 31). *Eugenics and statistics, discussing Karl Pearson and R. A. Fisher. meditations on inclusive statistics*. Nathaniel Joselson Meditations on Inclusive Statistics. https://njoselson.github.io/Fisher-Pearson/

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kai, S., Andres, J. M. L., Paquette, L., Baker, R. S., Molnar, K., Watkins, H., & Moore, M. (2017, June 25-28). *Predicting student retention from behavior in an online orientation course* [Paper presentation]. International Educational Data Mining Society's International Conference on Educational Data Mining, Wuhan, China.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January 7–10). Big data: Issues and challenges moving forward [Proceedings article]. 2013 46th Hawaii International Conference on System Sciences (HICSS), Wailea, HI. https://doi.org/10.1109/hicss.2013.645

Kaliyadan, F., & Kulkarni, V. (2019). Types of variables, descriptive statistics, and sample size. *Indian Dermatology Online Journal*, *10*(1), 82.

Karimshah, A., Wyder, M., Henman, P., Tay, D., Capelin, E., & Short, P. (2013). Overcoming adversity among low SES students: A study of strategies for retention. *Australian Universities' Review, 55*(2), 5-14.

King, N. (2020, June 29). Florida scientist says she was fired for not manipulating COVID-19 data. *National Public Radio*. https://www.npr.org/2020/06/29/884551391/florida-scientist-says-she-was-fired-for-not-manipulating-covid-19-data

Kitchin, R. (Ed.). (2014). Conceptualising data. In *The data revolution: Big data, open data, data infrastructures & their consequences* (pp. 1-26). Sage Publishing. https://doi.org/10.4135/9781473909472

Kitchin, R., & Lauriault, T. (2018). Towards critical data studies: Charting and unpacking data assemblages and their work. In J. Thatcher, J. Eckert, & A. Shears (Eds.), *Thinking big data in geography: New regimes, new research* (pp. 3-20). University of Nebraska Press. https://doi.org/10.2307/j.ctt21h4z6m.6

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G. M., & Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal of Clinical Epidemiology, 63*(7), 728-736. https://doi.org/10.1016/j.jclinepi.2009.08.028

Knox, J. (2017). Data power in education: Exploring critical awareness with the "Learning

Analytics Report Card". *Television & New Media*, *18*(8), 734-752. https://doi.org/10.1177/1527476417690029

Kuhn, T. (1970). *The structure of scientific revolution*. University of Chicago.

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior Research Methods, 49*(2), 457-470. https://doi.org/10.3758/s13428-016-0715-3

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology, 70*(4), 407. https://doi.org/10.4097/kjae.2017.70.4.407

Lavy, V., & Sand, E. (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases (No. w20909). National Bureau of Economic Research Working Paper Series. https://doi.org/10.3386/w20909

Levy, K. E., & Johns, D. M. (2016). When open data is a Trojan Horse: The weaponization of transparency in science and governance. *Big Data & Society, 3*(1), 2053951715621568. https://doi.org/10.1177/2053951715621568

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

Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Missing data. *Developmental Psychopathology,* 1-37. https://doi.org/10.1002/9781119125556.devpsy117

Lodder, P. (2013). To impute or not impute: That's the question. In G. J. Mellenbergh & H. J. Adèr (Eds.), *Advising on research methods: Selected topics* (pp. 1-7). Johannes van Kessel.

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal*, *296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Maestas, B., Stuntz, S., Applebee, J., Bentley, W., Breuker, J., Davenport, A., & Hoisington, A. (2020). A systematic approach to determine the amount of data required for asset management decisions. In J. Liyanage, J. Amadi-Echendu, J. Mathew (Eds.), *Engineering Assets and Public Infrastructures in the Age of Digitalization*. (pp.283-289). Springer. https://doi.org/10.1007/978-3-030-48021-9_32

Martin-Sanchez, F., & Verspoor, K. (2014). Big data in medicine is driving big changes. *Yearbook of Medical Informatics, 9*(1), 14. https://doi.org/10.15265/iy-2014-0020

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McMurtrie, B. (2018, May 25). Georgia State U. made its graduation rate jump. How? *The Chronicle of Higher Education*. https://www.chronicle.com/article/Georgia-State-U-Made-Its/243514

McNeely, C. L., & Hahm, J.-o. (2014). The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research*, *31*(4), 304-310. https://doi.org/10.1111/ropr.12082

Meeyai, S. (2016). Logistic regression with missing data: A comparison of handling methods and effects of percent missing values. *Journal of Traffic and Logistics Engineering*, *4*(2). https://doi.org/10.18178/jtle.4.2.128-134

Millea, M., Wills, R., Elder, A., & Molina, D. (2018). What matters in college student success? Determinants of college retention and graduation rates. Education, *138*(4), 309-322. https://www.ingentaconnect.com/contentone/prin/ed/2018/00000138/00000004/art00003

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society, 3*(2). https://doi.org/10.1177/2053951716679679

Moraes, M. C., Folkestad, J., & Birmingham, D. (2019, November). Critical lens in learning analytics research: A systematic literature review. Brazilian Symposium on Computers in Education [Symposium]. https://doi.org/10.5753/cbie.sbie.2019.1381

Murumba, J., & Micheni, E. (2017). Big data analytics in higher education: A review. *The International Journal of Engineering and Science*, *6*(06), 14-21. https://doi.org/10.9790/1813-0606021421

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures, 5*(4), 297-310. https://doi.org/10.1080/19312458.2011.624490

Nadworny, E. (2019, March 13). College completion rates are up, but the numbers will surprise you. *National Public Radio*. https://www.npr.org/2019/03/13/681621047/college-completion-rates-are-up-but-the-numbers-will-still-surprise-you

National Institute of Health. (2020). *Additional scoring guidance for research applications*. https://grants.nih.gov/grants/peer/guidelines_general/scoring_guidance_research.pdf

National Student Clearinghouse Research Center. (2018). *Snapshot report: Persistence-retention*. National Student Clearinghouse. https://nscresearchcenter.org/wp-content/uploads/SnapshotReport33.pdf

Newman, D. A., & Cottrell, J. M. (2015). *Missing data bias: Exactly how bad is pairwise deletion?* In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (p. 133–161). Routledge/Taylor & Francis Group.

Nyce, C. (2007). Predictive analytics white paper. American Institute for Charter Property Casualty Underwriter/Insurance Institute of America. http://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Omalu, B. I., DeKosky, S. T., Minster, R. L., Kamboh, M. I., Hamilton, R. L., & Wecht, C. H. (2005). Chronic traumatic encephalopathy in a National Football League player. Neurosurgery, *57*(1), 128-134. https://doi.org/10.1227/01.neu.0000163407.92769.ed

Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., & Casadesús-Fa, M. (2020). University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innovations in Education and Teaching International, 57*(1), 74-85. https://doi.org/10.1080/14703297.2018.1502090

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology and Society, *17*(4), 49-64. https://doi.org/10.1016/j.chb.2018.07.027

Pascarella, E. T., & Terenzini, P. T. (1998). Studying college students in the 21st century: Meeting new challenges. *The Review of Higher Education, 21*(2), 151–165.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840x.2017.1335568

Pearson, K. (1909). *The scope and importance to the state of the science of national eugenics* (2nd ed.). Dulau & Co.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, *74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. *Political Analysis*, *26*(4), 480-488. https://doi.org/10.1017/pan.2018.18

Pérez Huber, L., Vélez, V. N., & Solórzano, D. (2018). More than 'papelitos:'A QuantCrit counterstory to critique Latina/o degree value and occupational prestige. *Race Ethnicity and Education*, *21*(2), 208-230. https://doi.org/10.1080/13613324.2017.1377416

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, *16*(3), 9-20. https://doi.org/10.24059/olj.v16i3.267

Pilgrim, M. E., Folkestad, J. E., & Sencindiver, B. (2017, March 13-17). *Identifying non-regulators: Designing and deploying tools that detect self-regulation behaviors* (S. Shehata & J. P-L. Tan, Eds.). Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference, Simon Fraser University and SoLAR, Vancouver, Canada. https://solaresearch.org/wp-content/uploads/2017/02/Final-LAK17-Practitioner-Track-Proceedings.pdf

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components ofclassroom academic performance. *Journal of Educational Psychology*, *82*(1), 33.

Pitcan, M. (2016). Does data-driven learning improve equity? Data and Society. https://points.datasociety.net/does-data-driven-learning-improve-equity-8416ae173735

Poster, M. (2019). *Critical theory and poststructuralism: In search of a context*. Cornell University Press.

Premraj, D., Thompson, R., Hughes, L., & Adams, J. (2019). Key factors influencing retention rates among historically underrepresented student groups in STEM fields. *Journal of College Student Retention: Research, Theory & Practice, 23*(2), 457–478. https://doi.org/10.1177/1521025119848763

Raaijmakers, Q. A. W. (1999). Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59*(5), 725–748. https://doi.org/10.1177/00131649921970116

Rae, J. R., & Olson, K. R. (2018). Test–retest reliability and predictive validity of the Implicit Association Test in children. *Developmental psychology, 54*(2), 308-330. https://doi.org/10.1037/dev0000437

Rajuladevi, A. (2018). *A machine learning approach to predict first-year student retention rates at University of Nevada, Las Vegas* [Doctoral dissertation, University of Nevada Las Vegas]. Digital Scholarship@UNLV. https://digitalscholarship.unlv.edu/thesesdissertations/3315

Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments. *Journal of Student Affairs Research and Practice*, *40*(4), 704-723. https://doi.org/10.2202/1949-6605.5022

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology, 32*(2), 143.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry, 24*(4), 657-720.

Reyes, V. (2020). Ethnographic toolkit: Strategic positionality and researchers' visible and invisible tools in field research. *Ethnography, 21*(2), 220-240. https://doi.org/10.1177/1466138118805121

Rizkallah, E. G., & Seitz, V. (2017). Understanding student motivation: a key to retention in higher education. *Scientific Annals of Economics and Business*, *64*(1), 45-57. https://doi.org/10.1515/saeb-2017-0004

Roberts, A., & Kwon, R. (2017). Finance, inequality and the varieties of capitalism in post-industrial democracies. *Socio-Economic Review*, *15*(3), 511-538. https://doi.org/10.1093/ser/mwx021

Rubin, P. H. (1977). Why is the common law efficient? *The Journal of Legal Studies*, *6*(1), 51-63. https://doi.org/10.2139/ssrn.498645

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. *American Educational Research Journal, 56*(1), 178-203. https://doi.org/10.3102/0002831218798325

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing data. In *Secondary analysis of electronic health records* (pp. 143-162). Springer. https://doi.org/10.1007/978-3-319-43742-2_13

Samayoa, J. A., & Fazio, R. H. (2017). Who starts the wave? Let's not forget the role of the individual. *Psychological Inquiry*, *28*(4), 273-277. https://doi.org/10.1080/1047840X.2017.1373554

Savin-Baden, M., & Major C. (2013). *Qualitative research: The essential guide to theory andpractice*. Routledge.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3-15. https://doi.org/10.1177/096228029900800102

Schimmack, U. (2019). The implicit association test: a method in search of a construct. *Perspectives on Psychological Science*, 1745691619863798. https://doi.org/10.1177/1745691619863798

Schwartz, N. (2014). Cognition and communication: Judgmental biases, research methods and the logic of conversation. *Psychology Press*. https://doi.org/10.4324/9781315805887

Secules, S., McCall, C., Mejia, J. A., Beebe, C., Masters, A. S., L. Sánchez-Peña, M., & Svyantek, M. (2021). Positionality practices and dimensions of impact on equity

research: A collaborative inquiry and call to the community. *Journal of Engineering Education, 110*(1), 19-43. https://doi.org/10.1002/jee.20377

Seidman, A. (Ed.). (2005). *College student retention: Formula for student success*. Greenwood Publishing Group.

Shaw, R. M., Howe, J., Beazer, J., & Carr, T. (2020). Ethics and positionality in qualitative research with vulnerable and marginal groups. *Qualitative Research*, *20*(3), 277-293. https://doi.org/10.1177/1468794119841839

Shields, M. (2005). Information literacy, statistical literacy, data literacy. *IASSIST quarterly, 28*(2-3), 6-6. https://doi.org/10.29173/iq790

Sindhi, K., Parmar, D., & Gandhi, P. (2019). A study on benefits of big data for healthcare sector of India. In D. Mishra, XS. Yang, & A. Unal (Eds.), *Data science and big data analytics* (Vol. 16, pp. 239-246). Springer. https://doi.org/10.1007/978-981-10-7641-1_20

Sleek, S. (2018). The bias beneath: Two decades of measuring implicit associations. *APS Observer*, *31*(2).

Starkweather, J., & Herrington, R. (2018). *Data science and analytics*. University of North Texas. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/SPSS_SC/Module6/SPSS_M6_2.htm

Staw, B. (1981). Some judgments on the judgment calls approach. *American Behavioral Scientist, 25*(2), 225–232. https://doi.org/10.1177/000276428102500207

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Stone, D. A. (2013). *Policy paradox: The art of political decision making*. Norton.

Tinto, V. (1993). Building community. *Liberal education*, *79*(4), 16-21.

Tuttle, H. (2018). Facebook scandal raises data privacy concerns. *Risk Management, 65*(5), 6-9.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

United States Office of the President. (2014). Big data: Seizing opportunities, preserving values. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

University of Colorado Colorado Springs [UCCS]. (2020a). Fall Databook. https://ir.uccs.edu/sites/g/files/kjihxj1231/files/inline-files/FallDatabook_1.pdf

University of Colorado Colorado Springs [UCCS]. (2020b). Fall enrollment projection. https://ir.uccs.edu/sites/g/files/kjihxj1231/files/inline-files/Fall19toFall20_0.pdf

U.S. Department of Education. (2019). College scorecard. https://collegescorecard.ed.gov/data/

U.S. Department of Education. (2007). Parents' guide to the family educational rights and privacyact Rights regarding children's education records. https://www2.ed.gov/policy/gen/guid/fpco/brochures/parents.html

Van Velthoven, A., De Haan, J., & Sturm, J. E. (2019). Finance, income inequality and income redistribution. *Applied Economics Letters*, *26*(14), 1202-1209. https://doi.org/10.1080/13504851.2018.1542483

Vaske, J. J. (2008). *Survey research and analysis: Applications in parks*. Venture.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. Computers in Human Behavior, *89*, 98-110. https://doi.org/10.1016/j.chb.2018.07.027

Walker, S. S., & Schimmack, U. (2008). Validity of a Happiness Implicit Association Test as a measure of subjective wellbeing. *Journal of Research in Personality, 42*(2), 490–497. https://doi.org/10.1016/j.jrp.2007.07.005

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, *8*, 78-115. https://doi.org/10.5325/jinfopoli.8.2018.0078

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. Sage.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Sage.

Yoon, J., Jordon, J., & Van Der Schaar, M. (2018). *Gain: Missing data imputation using generative adversarial nets* [Conference proceedings]. The 35th International Conference on Machine Learning, Stockholm, Sweden. https://arxiv.org/pdf/1806.02920.pdf

Yousif, M. (2015). The rise of data capital. *IEEE Cloud Computing, 2*(2), 4. https://doi.org/10.1109/mcc.2015.39

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8*(2), 307-325. https://doi.org/10.6339/jds.2010.08(2).574

Yuan, Y. C. (2010). *Multiple imputation for missing data: Concepts and new development (Version 9.0)*. SAS Institute Inc. http://facweb.cdm.depaul.edu/sjost/csc423/documents/multipleimputation.pdf

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine, 4*(1). https://doi.org/10.3978/j.issn.2305-5839.2015.12.38

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Pant, H. A. (Eds.). (2018). *Assessment of learning outcomes in higher education: Cross-National comparisons and perspectives*. https://doi.org/10.1007/978-3-319-74338-7

Zuberi, T. (2001). *Thicker than blood: How racial statistics lie*. University of Minnesota Press.

APPENDIX A

Table A.1

Demographics of Listwise Deletion

Total financial aid	SAT score	ACT score	Retention	Gender	Ethnicity	Current semester enrolled	Student life club	First generation status
2500.00	1190	29	retained	F	Hispanic	Fall 2017	no	FirstGen
3062.50	1200	25	not	M	White	Fall 2019	no	Unknown
32282.20	1000	20	not	M	TwoPlus	Fall 2017	no	Not
5500.00	1190	24	retained	F	White	Fall 2019	no	Not
8000.00	1400	33	retained	F	White	Fall 2017	no	Not
5000.00	1090	21	retained	M	White	Fall 2017	no	Not
0.00	1290	21	retained	M	White	Fall 2017	yes	Not
36821.00	950	20	not	F	Hispanic	Fall 2017	no	Not
0.00	1070	23	retained	F	White	Fall 2017	yes	Not
9500.00	1060	23	retained	F	White	Fall 2019	no	FirstGen
2500.00	1330	26	not	M	White	Fall 2017	no	Not

0.00	980	19	not	M	White	Fall 2017	no	Not
2500.00	1310	28	retained	M	White	Fall 2019	no	Not
3000.00	1110	23	retained	M	White	Fall 2019	no	Not
28380.00	1170	27	retained	F	White	Fall 2019	no	Not
7244.00	1100	27	retained	M	White	Fall 2019	no	Not
10500.00	1220	27	retained	F	White	Fall 2019	yes	Not
9500.00	1030	23	retained	F	White	Fall 2019	no	Not
20823.00	1390	29	retained	F	White	Fall 2019	yes	Not
4000.00	970	26	not	F	Hispanic	Fall 2019	no	FirstGen
0.00	1330	29	retained	F	White	Fall 2019	no	Not
37812.00	1150	23	not	M	White	Fall 2019	no	Not
28380.00	1180	27	retained	F	White	Fall 2019	no	Not
16096.00	1140	21	retained	F	Hispanic	Fall 2019	no	FirstGen
2500.00	1110	25	retained	M	White	Fall 2019	no	Not
17000.00	1080	25	retained	F	TwoPlus	Fall 2019	yes	Not
13500.00	1470	31	retained	M	TwoPlus	Fall 2019	no	Not

4272.50	1060	23	not	F	Hispanic	Fall 2019	yes	Not
0.00	1090	22	retained	F	White	Fall 2019	no	Not
8500.00	1050	25	retained	M	Hispanic	Fall 2019	no	Not
8000.00	1220	27	retained	M	White	Fall 2019	no	Not
2500.00	1230	19	retained	M	Hispanic	Fall 2019	no	Not
22880.00	1250	23	not	M	White	Fall 2019	no	Not
23460.00	940	20	not	F	White	Fall 2019	no	Not
0.00	960	22	retained	M	White	Fall 2019	no	Not
11195.00	1060	25	retained	M	TwoPlus	Fall 2019	no	FirstGen