Using multiple imputation to mitigate the effects of low examinee motivation on estimates of student learning

Kelly J. Foelber
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/diss201019
Part of the Quantitative Psychology Commons

Recommended Citation
https://commons.lib.jmu.edu/diss201019/148

This Dissertation is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
Using Multiple Imputation to Mitigate the Effects of Low Examinee Motivation on Estimates of Student Learning

Kelly J. Foelber

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2017

FACULTY COMMITTEE:

Committee Chair: Dena Pastor, Ph.D.

Committee Members/Readers:

Deborah Bandalos, Ph.D.
Christine DeMars, Ph.D.
Daniel Jurich, Ph.D.
Acknowledgments

I would first like to acknowledge my family and friends for their unconditional support and encouragement throughout this process. In many ways, both small and not-so-small, you’ve all helped me see this project to completion. Thank you.

I would also like to acknowledge my dissertation committee for their insight and thoughtful feedback. Debbi, Christine, and Dan, I am grateful for the time and careful consideration you each put into helping me strengthen this dissertation.

Finally, I would like to acknowledge my advisor and chair of this dissertation, Dena. Dena, I am incredibly grateful for your patience, your perspective, and your positivity. It’s been my privilege to work with you and learn from you.
# Table of Contents

Acknowledgments .................................................................................................................. ii  

List of Tables ........................................................................................................................... vii  

List of Figures ............................................................................................................................ ix  

Abstract ..................................................................................................................................... x  

Introduction ............................................................................................................................... 1  

Literature Review ....................................................................................................................... 11  

Introduction ............................................................................................................................... 11  

Addressing Examinee Motivation ............................................................................................ 12  

Measuring examinee motivation ............................................................................................... 12  

Self-report measures ............................................................................................................... 12  

Item response time .................................................................................................................. 14  

Summary ................................................................................................................................... 19  

Post-hoc approaches for addressing examinee motivation ....................................................... 19  

Motivation filtering ................................................................................................................... 20  

Rapid-response filtering ........................................................................................................... 23  

Effort-moderated IRT model .................................................................................................... 23  

Missing Data ............................................................................................................................. 24  

Missing data theory ................................................................................................................ 25  

Missingness mechanisms ......................................................................................................... 26
Data Collection .....................................................................................................................56
Examinees ..............................................................................................................................57
Measures .................................................................................................................................58
The Natural World Test version 9 ........................................................................................58
Examinee effort .......................................................................................................................59
Auxiliary variables .................................................................................................................61
Data Analysis ........................................................................................................................63
Post-hoc methods for addressing examinee motivation .........................................................63
Assessing the plausibility of the MCAR assumption .............................................................66
Substantive analyses ..............................................................................................................67
Software ................................................................................................................................70
Results ....................................................................................................................................71
Introduction ............................................................................................................................71
Description of Data ................................................................................................................71
Observed NW-9 responses ......................................................................................................71
Examinee effort .......................................................................................................................72
Auxiliary variables .................................................................................................................74
Applying Post-Hoc Methods .................................................................................................74
Assessing the Plausibility of the MCAR Assumption .............................................................76
Substantive Analyses ..............................................................................................................78
Mixed ANOVA results........................................................................................................79

Comparing MI results by number of imputed datasets..............................................81

Discussion.......................................................................................................................83

Introduction..................................................................................................................83

Purposes of this Dissertation.........................................................................................83

Summary of Results by Research Question...............................................................86

Research question one..................................................................................................86

Research question two..................................................................................................89

Research question three..............................................................................................91

Research question four...............................................................................................92

Recommendations for Researchers and Practitioners.................................................94

Primary Limitations of this Study...............................................................................97

Future Research..........................................................................................................100

Summary and Conclusion.........................................................................................101

Appendix A: Obtaining parameter estimates and multiparameter significance tests
in SAS..............................................................................................................................103

Appendix B: R syntax for setting CUMP thresholds.................................................105

Appendix C: Blimp syntax for imputing data.............................................................107

Appendix D: SAS syntax............................................................................................109

References..................................................................................................................118
List of Tables

Table 1: Response Accuracy by Response Time for Item 8 of the NW-9 Pre-Test Administration

Table 2: Example dataset

Table 3: Example Dataset for Deletion

Table 4: Sample Descriptive Information

Table 5: Thresholds and Threshold-Setting Methods for Each Item at Each Time Point

Table 6: Distribution of Number of Cluster 3 Courses Completed by Post-Test

Table 7: Amount of Missing Data Within Examinees

Table 8: Results of Independent-Groups T-Tests Between Low- and Adequately-Motivated Examinees on Auxiliary Variables

Table 9: Descriptive Statistics for NW-9 Scores: Observed Data (Unaltered)

Table 10: Descriptive Statistics for NW-9 Scores: Motivation Filtered Data

Table 11: Descriptive Statistics for NW-9 Scores: MI Data (m=100)

Table 12: Descriptive Statistics for NW-9 Difference Scores by Post-Hoc Method

Table 13: Descriptive Statistics for NW-9 Difference Scores by Number of Cluster Three Courses and Post-Hoc Method

Table 14: Mixed ANOVA Unstandardized Parameter Estimates and Standard Errors

Table 15: Tests of Mixed ANOVA Main Effects and Interaction
Table 16: Pooled Means and Standard Errors For NW-9 Total Scores by Number of Imputations .................................................................142

Table 17: Pooled Mixed ANOVA Parameter Estimates and Standard Errors by Number of Imputations ........................................................................................................143

Table 18: Tests of Mixed ANOVA Main Effects and Interaction by Number of Imputations ........................................................................................................144

Table 19: Between- and Within-Imputation Variance Estimates by Number of Imputations ........................................................................................................145

Table 20: Between- and Within-Imputation Variance Estimates by Number of Imputations ........................................................................................................146
List of Figures

Figure 1: The distribution of response times for item 5 of the NW-9 post-test administration .................................................................148

Figure 2: The cumulative proportion correct function (solid line) and chance correct function (dashed line) for item 8 of the NW-9 pre-test administration ..........149

Figure 3: The three mechanisms of missing data.........................................................150

Figure 4: Median item response times for each item of the NW-9..................................151

Figure 5: Item response time thresholds for each item of the NW-9...............................152

Figure 6: Overlapping distributions of examinees’ RTE scores.......................................153

Figure 7: Average NW-9 scores by number of Cluster 3 courses completed at post-test, time, and post-hoc method.................................................................154

Figure 8: Average NW-9 scores by number of Cluster 3 courses completed at post-test, time, and post-hoc method.................................................................155
Abstract

In higher education, we often collect data in order to make inferences about student learning, and ultimately, in order to make evidence-based changes to try to improve student learning. The validity of the inferences we make, however, depends on the quality of the data we collect. Low examinee motivation compromises these inferences; research suggests that low examinee motivation can lead to inaccurate estimates of examinees’ ability (e.g., Wise & DeMars, 2005). To obtain data that better represent what students know, think, and can do, practitioners must consider, and attempt to negate the effects of, low examinee motivation. The primary purpose of this dissertation was to compare three methods for addressing low examinee motivation following data collection (i.e., “post-hoc” methods): (1) leaving the data as they were observed (leaving rapid responses intact), (2) motivation filtering (listwise deleting examinees with more than an acceptable amount of rapid responses), and (3) using multiple imputation with auxiliary variables to impute plausible solution-behavior responses in place of rapid responses. The data analyzed in this study came from the Natural World Test (NW-9; Sundre, 2008), which was administered to James Madison University students before and after completing coursework designed to improve their quantitative and scientific reasoning skills (and thus their NW-9 scores). After applying the three methods, mixed ANOVAs were performed to investigate the main effects of time and number of courses completed, and their interaction, on examinees’ scores. These analyses aligned with the following overarching question: Do the inferences we make about student learning depend on the post-hoc method used to address low examinee effort? Of the three methods, motivation filtering produced the highest estimates of examinee ability. Leaving the data as they
were observed produced the lowest estimates. Multiple imputation produced estimates between those from the previous two methods. Although the estimates differed by post-hoc method, the same substantive conclusions were reached. For this study, regardless of post-hoc approach, we concluded that examinees’ scientific and quantitative reasoning abilities changed over time, and that examinees who completed more relevant courses did not change significantly more than examinees who completed fewer relevant courses.
CHAPTER ONE

Introduction

In higher education, we often collect data in order to make inferences about student learning, and ultimately, in order to make evidence-based changes to programming or curricula to try to improve student learning. We use students’ assessment scores to make inferences about what they know, think, or can do. The inferences we make, however, depend heavily on the quality of the data we collect. If the quality of the data is compromised, the inferences we make about students’ abilities based on test performance may not be valid. Validity refers to “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA, & NCME, 2014, p. 11). When the test scores we observe represent or include information about something other than the construct of interest (i.e., include construct-irrelevant variance), validity is threatened. One cause of construct-irrelevant variance is low examinee motivation.

Examinee motivation refers to the extent to which students are motivated to complete a given assessment in accordance with their ability. Low examinee motivation is particularly problematic because it threatens the alignment between students’ observed test scores and their abilities. For example, an unmotivated examinee may respond carelessly to an assessment, resulting in a low overall score. This observed overall score is thus not likely indicative of this examinee’s ability (because the examinee did not demonstrate his or her ability when responding). However, all we have to make inferences from is the examinee’s observed score. Low examinee motivation compromises the inferences we make based on the students’ observed test scores because
we often assume that these observed scores reflect students’ abilities. Examinee motivation is especially concerning when assessments have low stakes for examinees (Wise & Smith, 2016).

At James Madison University (JMU), we collect large amounts of data regarding student learning and development through low-stakes assessments. That is, much of the data we collect come from assessments whose results do not have direct personal consequences for students. Thus, unfortunately, some students are insufficiently motivated to complete the assessments in accordance with their abilities. However, we still use these data to make inferences about students’ learning and development pertaining to general education, student affairs programs, and other campus-wide initiatives. Ultimately, the validity of these inferences depends on the motivation of the examinees.

Research suggests that low examinee motivation can lead to inaccurate estimates of examinee ability (Wise & DeMars, 2005), relationships between ability and other variables (Wise, 2009), and growth (Finney, Sundre, Swain, & Williams, 2016; Wise & DeMars, 2010). This means that not only can low examinee motivation affect our perceptions and conclusions about what examinees know, think, or can do, but also about what predicts or relates to their ability, as well as about the change in their ability over time. Thus, any interventions or changes in curricula made based on these inferences may be misaligned with what students actually need. For example, if low examinee motivation facilitates a relationship between ability and gender, with results indicating that males are of lower ability than females, practitioners may use this information to implement interventions targeted towards males (e.g., increasing the variety of the content used in
examples or assignments). However, the resources spent on such interventions may be for naught if this relationship exists solely as a function of motivation (if males were less motivated than females to perform in accordance with their abilities on the assessment).

As another example, if low examinee motivation attenuates estimates of growth in ability, practitioners may use this information to conclude that the intervention or programming that occurred between the assessments was ineffective. In higher education, such low ‘value-added’ estimates may lead to the conclusion that the educational programming is not helping students learn. Any subsequent changes made to the educational programming constitute potential misuses of resources that could be better spent on more impactful changes.

Inaccurate estimates of change over time can instigate a variety of negative consequences for institutions, in particular, because they rely on value-added estimates for not only informing changes intended to improve student learning, but also for accountability and accreditation purposes (see Finney et al., 2016). That is, estimates of change over time in student learning not only impact the resources allocated within institutions (e.g., resources spent on targeted interventions to improve student learning), but also the resources allocated to institutions (e.g., federal funding through accreditation). Given the consequences associated with test scores, accurate estimates of growth in student learning are of paramount importance.

In order to achieve data that better represent what students know, think, and can do, practitioners must consider, and attempt to negate the effects of, low examinee motivation. It is our responsibility to measure and account for examinee motivation,
particularly when the assessments are low-stakes for students (AERA, APA, & NCME, 2014; Wise, 2009).

Perhaps the two most common methods for measuring examinee motivation are via self-report measures, and via time spent responding to assessment items (i.e., the time that elapses from when an examinee is first presented with an item to the time when he or she completes the item). At JMU, we often use both. That is, not only do we ask students to report how much effort they expended completing the assessments, but we also often record item response time information for assessments that are administered electronically. These two types of measures have been found to provide similar results with regard to examinee motivation (Swerdzewski, Harmes, & Finney, 2011).

Although collecting information about examinee motivation helps researchers better understand the experience of the examinees, and thus the quality of the assessment data, such information should subsequently be used to make changes to improve the quality of the data. Examinee motivation data can, and should, be used in two ways. First, if a similar assessment data collection process is planned to occur in the future, examinee motivation data should be used to inform changes intended to improve the motivational climate of the testing environment. Certainly the best way to mitigate the negative effects of low examinee motivation on data quality is to motivate the examinees. However, when efforts to motivate are not effective for all examinees, post-hoc (post data collection) methods should be considered. This is the second way in which examinee motivation data can, and should, be used: to inform and interpret data analyses.

Currently, one of the most popularly used and recommended methods for handling data from low-motivated examinees is motivation filtering. Motivation filtering
(Sundre & Wise, 2003) occurs when the data from low-motivated examinees are removed from the dataset (and all substantive analyses). Substantive analyses are therefore based only on data from examinees who are believed to have put forth effort into their responses. Thus, the analyses are based only on data that are most likely to reflect examinees’ actual abilities. Research on motivation filtering generally agrees that the technique results in higher aggregate scores, lower variance of scores, and higher convergent validity of scores, compared to results obtained without applying motivation filtering (i.e., compared to using all of the observed data; Wise & DeMars, 2010). Ultimately, Wise and DeMars (2010) recommend that “measurement practitioners routinely apply motivation filtering whenever the data from low-stakes tests are used to support program decisions” (p. 27).

Blind application of motivation filtering, however, is not recommended. Practitioners must first consider whether or not their data meet the assumptions of motivation filtering. For example, an important assumption of motivation filtering is that examinee effort is not related to examinee ability. That is, by applying motivation filtering we are assuming that there is no difference in true ability levels between those who put forth effort (whose scores are retained and analyzed) and those who did not put forth effort (whose scores are filtered and removed from the dataset). Unfortunately, research investigating the relationship between examinee effort and ability has produced mixed results. Although many studies have not found support for a relationship between effort and ability (e.g., Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005), some have found support for such a relationship (e.g., Rios, Guo, Mao, & Liu,
2017; Wise, Pastor, & Kong, 2009). Even the possibility of a relationship between effort and ability should caution researchers.

If researchers suspect a relationship between examinee ability and examinee effort, and thus may not wish to use motivation filtering, what then are they to do in order to mitigate the effects of motivation on the inferences they make from their data? One possible solution, the subject of the current study, requires researchers to think about their observed data as having a missing data problem.

Reframing the problem of examinee motivation as a problem of missing data should not be a far stretch for researchers familiar with motivation filtering. By removing data from low-motivated examinees (i.e., motivation filtering), we are creating missingness. We are saying that the low-motivated examinees’ observed data are so unrepresentative of the examinees’ abilities that these data may as well be missing, and so we remove their data from the dataset (i.e., we set these values to missing). We can even think of these data as red herrings: the presence of these observed untrustworthy responses distracts us from the important realization that they are not providing us with the information we want—such information is missing. By removing data at the examinee level, we are thus applying listwise deletion. Therefore, motivation filtering can be reframed as a listwise deletion technique.

Like motivation filtering, listwise deletion relies on the assumption that the examinees whose data are being removed are not different from the examinees whose data are retained to yield accurate results. More formally, listwise deletion assumes that the data are missing completely at random (MCAR; Enders, 2010). This means that missingness (i.e., whether or not a value is missing) is neither related to the unobserved,
true values of the variables with missing data, nor is it related to any other measured variable in the dataset. In the context of motivation filtering, this means that examinee effort can neither be related to the unobserved, effortful values of the variables for which the examinee did not put forth effort in responding to in reality, nor can it be related to any other measured variable in the dataset (e.g., demographic variables, scores on other assessments). If this assumption is untenable, listwise deletion, and thus motivation filtering, can result in inaccurate parameter estimates (e.g., estimates of examinee ability). This brings us back to our earlier question: If researchers suspect a relationship between examinee ability and examinee effort, and thus may not wish to use motivation filtering, what then are they to do?

Although all of the traditional missing data techniques, including listwise deletion, require the strict MCAR assumption, there are two ‘modern’ missing data techniques that do not require this assumption. Instead, both full information maximum likelihood (FIML) estimation and multiple imputation (MI) assume that the data are missing at random (MAR); that is, missingness is not related to the unobserved, true values after controlling for other measured variables in the dataset. For researchers concerned about a relationship between low examinee effort (missingness) and examinees’ unobserved, true responses, the MAR assumption can hold if other measured variables are able to account for this relationship. For example, if effort (missingness) was related to the ability measured by the variable with missingness, a second variable measuring the same ability as the variable with missingness could be used to control for this relationship. Such variables are called auxiliary variables. That is, we would not expect there to be a relationship between effort (missingness) and the variable with
missingness, once controlling for the auxiliary variable. But if examinee effort is an issue, how are we to obtain an auxiliary variable measuring the same ability?

Unlike motivation filtering (listwise deletion), both FIML and MI can be applied at the item level. This means that rather than filtering (deleting) low-motivated examinees’ entire response strings, we can filter out only the specific responses associated with low examinee effort (i.e., set item-level responses to missing if associated with low effort). Working at the item level is advantageous in the context of examinee motivation because it means that entire strings of examinee data do not need to be discarded if the examinee displays low motivation on only some of the items, protecting your sample size from unnecessary depletion. If only some of an examinee’s item responses are untrustworthy, the other trustworthy responses can be retained and analyzed. Importantly, if some effortful responses are retained, and the items associated with these responses measure the same ability as any of the items with non-effortful responses (e.g., if the scale is unidimensional or the items are on the same subscale), these effortful responses make ideal auxiliary variables. Recall that the MAR assumption, made by both FIML and MI, requires missingness and the ability measured by the variable with missingness to be unrelated once controlling for other measured variables in the dataset. So, if we have a unidimensional measure with non-effortful responses (missingness) for some items, but not all items, we can use the effortful item responses as auxiliary variables, effectively controlling for a relationship between effort (missingness) and ability, and likely satisfying the MAR assumption. This is a major advantage of using the ‘modern’ missing data methods (at the item level). Again, though, using one of these
approaches requires researchers to first conceptualize their examinee effort problem as a missing data problem, which aligns with one of the purposes of this dissertation.

This dissertation has two primary purposes, one didactic in nature and one empirical in nature. The first purpose is to reframe the problem of examinee effort as a problem of missing data, and consider the manner in which existing post-hoc methods for addressing examinee effort align with missing data techniques. For instance, even though motivation filtering is listwise deletion, it is rare for it to be described as such or for the assumptions of listwise deletion (namely MCAR) to be fully considered. By reframing examinee effort as a missing data problem in the second chapter of this dissertation, I hope to illuminate the options researchers have for appropriately analyzing their data, as well as to draw attention to the assumptions made by various post-hoc methods (e.g., motivation filtering).

The second purpose, which is empirical in nature, is to further investigate the use of the ‘modern’ missing data method of MI, applied at the item level, in mitigating the effects of low examinee effort on estimates of examinee ability. Although there are many advantages to using MI, only one study thus far has considered its use in the context of low examinee effort (Koepfler, Jurich, & DeMars, 2011). Conceptually, MI works by replacing single missing data points with multiple, plausible predicted values. For this study, prior to applying multiple imputation we first imposed missingness in our data for all non-effortful, item-level responses. These responses were identified based on examinees’ item response times (item responses were categorized as effortful or not, based on how much time examinees used to respond). Then, we used MI to predict and impute values that may be better aligned with examinees’ true ability, compared to the
observed, non-effortful responses. Importantly, these imputed values were predicted using all effortful item-level responses as auxiliary variables. That is, we used the item responses for which examinees put forth effort (which we believe are indicative of their true ability), to help predict and impute responses for items on which the examinees did not put forth effort.

In order to investigate the utility of this method, this study used real data to compare the estimates of student growth obtained from applying three techniques: MI, motivation filtering, and analyzing the data as they were observed (unaltered).

Specifically, data from the Natural World Test version 9 (NW-9; Sundre, 2008) were collected from undergraduate students at JMU both before and after completing general education coursework designed to improve students’ quantitative and scientific reasoning skills, and thus to improve students’ NW-9 scores. These data were collected under low-stakes testing conditions and contain non-effortful responses from students. To address the non-effortful responding, pre-post growth estimates were compared after applying two post host methods, namely MI and motivation filtering. These pre-post growth estimates were also compared to those obtained by leaving the data as they were observed, replete with non-effortful responses. These analyses were used to address a total of four research questions, at the heart of which is the following overarching question: Do the inferences we make about student learning depend on the post-hoc method used to address low examinee effort?
CHAPTER TWO

Literature Review

Introduction

The purpose of this chapter is to review literature that serves as the foundation for this study. Three major topics are covered. The chapter begins with the topic of current post-hoc methods for addressing examinee motivation in low-stakes testing environments. I first discuss how examinee motivation is measured, and then describe existing approaches for addressing low examinee motivation. The second major section of this chapter provides an overview of the topic of missing data, including missing data theory and methods for addressing missing data. Because one of the purposes of this study is to reframe the problem of low examinee motivation as a problem of missing data, this second section is included to prepare readers for this connection by providing relevant background information from the missing data literature. The third and final section of this chapter serves to combine ideas from both of the previous topics. Specifically, I explicitly connect existing post-hoc methods for addressing examinee motivation to existing missing data methods (e.g., motivation filtering to listwise deletion). Further, I discuss literature that has actually applied one of these post-hoc missing data methods in the context of examinee motivation, even if the researchers did not describe it as such. The importance of this third section is underscored by the fact that this connection, between post-hoc examinee motivation methods and missing data methods, is only acknowledged by a single study reviewed here. I conclude this chapter by summarizing the purpose of the empirical study and describing the four research questions I intend to address.
Addressing Examinee Motivation

One of the purposes of this study is to reframe current post-hoc methods for addressing examinee motivation as missing data methods. In order to fulfill this purpose, the current post-hoc methods must first be understood. Because all post-hoc methods rely on the pre-identification of non-effortful responses, this process is first discussed. Specifically, this section begins with a description of how self-report scales and item response times are used to measure examinee motivation and identify non-effortful responses. Because item response times are used as measures of effort in the present study, particular attention is paid to how response time thresholds for differentiating effortful vs. non-effortful responses are determined using the cumulative proportion (CUMP) method (Guo et al., 2016). The section concludes with a description of current methods used to address motivation, with particular attention paid to three post-hoc methods: motivation filtering, rapid-response filtering, and the effort-moderated item response theory (IRT) model.

Measuring examinee motivation. Measuring examinee effort is an important first step in the process of better understanding examinees’ responses to measures of interest and thus examinee ability. When we consider measuring examinee test-taking effort, we have two primary options: self-report measures and item response times.

Self-report measures. Although there are various ways to quantify examinee motivation, self-report is perhaps the most common method. Popular self-report measures include the Student Opinion Scale (SOS; Sundre & Moore, 2002), the Test-Taking Motivation Questionnaire (TTMQ; Eklöf, 2006; Knekta & Eklöf, 2015) and the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990).
Although easy to use, self-report measures are not without their disadvantages. For instance, self-reported responses are filtered through the lenses of the examinees, often after completing the measure or measures of interest. Many researchers note that examinees may be prone to letting their perceived performance affect how much effort they report having put forth (e.g., Wise & Kong, 2005; Wise & Smith, 2016). For example, if examinees believed they performed poorly on the measure in question they may be more likely to attribute their poor performance to a lack of effort, rather than a lack of proficiency, and self-report that they had put forth little effort into their responses.

Another potential issue with self-report measures is that they require examinee effort. In order to make accurate inferences about examinees’ exam scores, we rely on examinee motivation information; this information itself, however, relies on examinees being motivated enough to respond appropriately to the measure of motivation. If examinees have little motivation to respond to the assessments of interest in accordance with their ability, it is quite possible that they also have little motivation to accurately respond to the self-report measure of motivation itself. In such cases, the information used to determine the trustworthiness of the assessment data may also in fact be untrustworthy.

A final potential issue with self-report measures of examinee motivation given at the end of testing is that they elicit examinees’ motivation in an overall sense (i.e., examinees’ motivation generalized across the test-taking session). Examinees are thus unable to report different levels of motivation that may have occurred throughout the session; they are simply asked about their test-taking motivation in general. Unfortunately, summarizing examinee motivation ignores valuable information regarding
the potential variation of motivation both within and between measures. Differences in effort throughout testing-sessions have been documented across various research efforts (e.g., Barry & Finney, 2016; Barry, Horst, Finney, Brown, & Kopp, 2010; DeMars, 2007; Horst, 2010; Penk & Richter, 2017; Wise & Kong, 2005; Wise, 2006).

**Item response time.** In addition to self-report measures, researchers have quantified examinee motivation using the time examinees spend completing a measure, or more specifically, the time they spend on each item (i.e., the time between the initial presentation of the item and when they provide or submit a response). Item response time measures present two primary advantages over self-report measures: they more covertly measure examinee motivation (the examinee is unaware that motivation is being measured), and they are capable of measuring fluctuations in motivation both within, at the item level, and across measures. Schnipke and Scrams (1997) theorized that examinees exhibit one of two response strategies when presented with an item, coining the terms ‘solution behavior’ and ‘rapid-guessing behavior’. According to the authors, rapid-guessing behavior occurs when an examinee responds to an item in so little time that he or she could not have “read and fully consider[ed]” the item (Schnipke & Scrams, 1997, p. 214). Subsequently, a time threshold must be set in order to determine whether an examinee exhibited rapid-guessing behavior for a given item. Solution behavior, on the other hand, occurs when an examinee’s item response time is greater than (or equal to) the determined time threshold. It is thus assumed that the examinee actively tried to identify the solution to the given item (Schnipke & Scrams, 1997). Although Schnipke and Scrams’ (1997) research was conducted in the context of speeded tests, the dual response strategy theory has also been supported in the context of low-stakes testing.
Specifically, Wise and Kong (2005) found evidence to support their hypothesis that, in the context of low-stakes testing, examinees exhibit either solution behavior or rapid-guessing behavior for each item on a measure.

Selection of the time thresholds used to determine examinees’ response strategies is completed on an item-by-item basis, and can be accomplished using various methods (see Guo et al., 2016). Time threshold selection is often informed by information such as item characteristics (e.g., number of characters, whether or not examinees must reference other materials), and/or observed response times for the examinees. Thus, thresholds are typically determined after data have been collected. One of the simplest threshold setting techniques is to select a standard time for all items (e.g., 3 seconds; Kong, Wise, & Bhola, 2007). Another fairly simple technique is visual inspection (VI; Wise, 2006). VI is performed by visually inspecting histograms of observed response times for each item, which are assumed to portray two populations of responders: rapid responders and non-rapid responders. These two populations are often indicated by a bimodal response time histogram. The VI method defines an item’s threshold as the response time associated with the location at which it appears the rapid responders’ distribution ends (where the smaller mode’s distribution ends). For example, in Figure 1 we see the (bimodal) distribution of responses times for an item. According to the VI method, this item’s threshold would likely be set just shy of 10 seconds because this is where the first distribution appears to end.

A newer, more complex method for determining item response time thresholds is the cumulative proportion (CUMP) method (Guo et al., 2016), which incorporates both response time and response accuracy information into the threshold selection process.
Specifically, an item’s response time threshold is identified as the time corresponding to the intersection at which the cumulative proportion correct function exceeds the chance correct function for the item, as depicted in Table 1 and Figure 2. To accomplish this, first the observed proportion of examinees correctly responding to the item is accumulated as a function of item response times, creating a (cumulative) function across time. Column 6 in Table 1 contains the cumulative proportion correct function for item 8 from the pre-test administration of the NW-9. The interpretation for row 5 of Table 1, as an example, is as follows: out of the 10 examinees who responded to this item in 15 seconds or less, 2 out of 10 (or 20%) provided the correct response. Then, the probability of correctly responding to the item by chance is determined (e.g., 1/number of response options), creating a constant function across time. For item 8 there were 3 response options, and so column 7 displays the constant 1/3 (approximately 0.3333). The item’s threshold is defined as the point at which the first function (cumulative proportion correct) exceeds the second function (chance correct). That is, the response time used to distinguish between rapid responses and responses provided under solution behavior is determined as the response time for which examinees who responded in that time or less demonstrate a chance level of responding correctly. For item 8 we see that this intersection occurs between 21 and 22 seconds. That is, less than 33% of examinees who responded in 21 seconds or less responded correctly, and more than 33% of examinees who responded in 22 seconds or less responded correctly. Because response times are only recorded in whole numbers here, and in this study, the threshold for this item would be set at 22 seconds. The cumulative proportion correct function and chance correct function for item 8 are shown in Figure 2.
The solid line in Figure 2 displays the cumulative proportion correct function for item 8 for the pre-test administration of the NW9 (column 6 of Table 1). For each item response time (horizontal axis), the cumulative proportion correct function (solid line) indicates the proportion of examinees who selected the correct response for this item (vertical axis) in that amount of time or less. For example, the cumulative proportion correct function value at 15 seconds, 0.20, indicates that 20% of examinees who spent 15 seconds or less on this item selected the correct response. The dashed line in Figure 2 displays the chance correct function. Because there were 3 response options for this item, the chance correct function is constant at 1/3, or about 0.33. The CUMP method threshold for this item is indicated by the intersection of the solid and dashed lines, between 21 and 22 seconds. Again, because response times are only recorded in whole numbers here, and in this study, the threshold for this item would be set at 22 seconds.

Although it is new, and thus has seen little implementation (Guo et al., 2016; Rios et al., 2017), the CUMP method has certain advantages over other threshold-setting methods. One such advantage is that it incorporates observed response accuracy information, in addition to observed response time information, into the threshold-setting process. Thus, we can take examinee performance into consideration when making threshold determinations. This means that for those examinees who spend little time responding to an item we can also know the rate at which they correctly respond. We assume that rapid responses, on average, produce correct responses at the same rate as we would expect by chance (e.g., by guessing), and so we determine the response time associated with chance-level correct responding for the collection of examinees responding at that time or less. Another advantage of the CUMP method is that it uses
cumulative response accuracy information, rather than response accuracy frequencies which are used by the methods that served as the foundation for the CUMP method (see Guo et al., 2016). The advantage of using the cumulative function is that it addresses issues of sparseness of response accuracy data across the response time spectrum (if there are few examinees at some response time points).

In comparing the performance of the CUMP method to two other threshold setting methods (the quantile and visual inspection methods), Guo et al. (2016) concluded that the CUMP method performed as well as, or better than, the others with regard to IRT model fit, item parameter estimation, and score estimation. Although the VI method performed similarly, the authors noted that the CUMP method was easier to implement. To reiterate though, the CUMP method is new, and thus it has yet to be subjected to thorough investigation.

Regardless of the method used, once thresholds are determined for each item they are then compared to examinee response times for each item. Examinee responses that have response times greater than or equal to the designated threshold for a given item are categorized as solution behavior. Examinee responses that have response times less than the designated threshold for a given item are categorized as rapid-guessing behavior. For the item used as the example in Table 1 and Figure 2, examinees’ responses associated with response times greater than or equal to 22 seconds (the response time immediately following the intersection of the two functions) would be categorized as solution behavior responses, and those associated with response times less than 22 seconds would be categorized as rapid responses. Response behavior is commonly denoted using a single binary variable that refers to the presence or absence of solution behavior.
Although each item is evaluated for each examinee with regard to whether or not solution behavior was exhibited, an overall indication of effort can also be examined and used to indicate examinee motivation. For this purpose, Wise and Kong (2005) presented the response time effort (RTE) index. An examinee’s RTE is simply the proportion of items on a measure for which the examinee exhibited solution behavior. For example, if an examinee exhibited solution behavior on 50 items of a 75-item measure (and thus rapid-guessing behavior on 25 items), this examinee’s RTE score for this measure would be approximately 0.67 (50/75 ≈ 0.67), meaning solution behavior was exhibited for approximately 67% of the items. RTE scores range from zero to one, with higher scores indicating a greater proportion of test items for which an examinee exhibited solution behavior.

**Summary.** Measuring examinee effort is an important first step in the use of post-hoc methods for addressing examinee motivation. Test-level indicators of effort that represent the typical effort level of the examinee on the test can be acquired using either self-report methods or RTE scores. The only item-level indicators of effort that currently exist are based on item response time. The distinction between item- and test-level indicators of effort is relevant to the following discussion of post-hoc methods for addressing examinee motivation, where some methods (e.g., motivation filtering) use test-level measures of effort, and others require the use of item-level measures.

**Post-hoc approaches for addressing examinee motivation.** Approaches for addressing examinee motivation in low-stakes testing either attempt to address motivation a priori (i.e., prior to the completion of data collection) or post-hoc (i.e., after the data have all been collected). Although in an ideal scenario researchers would be able
to successfully apply a priori methods to adequately motivate all examinees, we must realize that in reality, no matter what we do, we may not be able to adequately motivate all examinees to complete low-stakes assessments to the best of their ability. Hence the necessity of post-hoc methods.

Post-hoc methods for addressing examinee motivation typically involve statistical interventions, although, one method for addressing examinee motivation simply involves issuing a disclaimer along with low-stakes assessment data. That is, the data are left intact, yet score users are provided with information acknowledging the levels of examinee motivation associated with the data. One of the issues with this method, though, is that it assumes score users will be able to incorporate the examinee motivation information to make valid inferences from the data; if not, this is essentially the same as assuming the low-stakes data are truly reflective of students’ abilities. In the sections that follow, I describe three post-hoc methods for addressing examinee motivation that involve statistical interventions.

**Motivation filtering.** Perhaps the most commonly used post-hoc method is motivation filtering (Sundre & Wise, 2003). The general idea of motivation filtering is that we can improve the validity of the inferences we make about our data if we remove, or filter out, data from unmotivated examinees. In applying motivation filtering we are essentially removing data deemed untrustworthy and subsequently analyzing only the remaining data, which we believe are truly reflective of examinees’ knowledge and abilities. Motivation filtering occurs at the examinee level; examinees are classified as motivated or unmotivated and only those in the former category are retained for analyses.
Motivation filtering relies on two primary assumptions. First, that we are able to accurately measure examinee motivation (i.e., that our examinee motivation scores are reliable and neither over- nor under-represent the construct of examinee motivation). Second, that examinee motivation is unrelated to true examinee ability. If motivation and ability were positively related, motivation filtering would effectively be removing the examinees of lower ability. This in turn would positively bias estimates of group ability. If motivation and ability were unrelated, filtering out low motivated examinees should not impact estimates of group ability.

In their introduction and initial study of motivation filtering, Sundre and Wise (2003) filtered examinees based on both self-reported motivation scores and patterns of responding to the self-report measure (e.g., completing only a few items, providing contradictory responses). They found that as more examinees were filtered from the dataset (i.e., as more stringent motivation filters were applied) test performance estimates increased, and the correlation between test performance and SAT total (a proxy for ability) increased. These results align with the idea that test performance is related to motivation. The authors also found that as more examinees were filtered, the correlation between self-reported motivation and SAT total (a proxy for ability) remained near zero, aligning with research that does not support a relationship between motivation and ability. Ultimately, Sundre and Wise's (2003) results supported their hypothesis that using motivation filtering to remove untrustworthy data from low-motivated examinees served to reduce “distortions in our assessment of the proficiency levels of a group of examinees” (p. 13). The authors encouraged the use of motivation filtering in estimating group proficiency based on data from low-stakes assessments.
Although Sundre and Wise (2003) introduced motivation filtering using self-reported motivation scores, other researchers have applied motivation filtering based on other measures of examinee motivation, namely, RTE scores (e.g., Kong et al., 2007; Rios, Liu, & Bridgeman, 2014; Swerdzewski et al., 2011). Importantly, research has also been conducted to compare the results of motivation filtering using self-reported effort and that using RTE scores. Wise and Kong (2005) and Swerdzewski et al. (2011) found that both methods produced similar aggregate test scores, although self-reported effort tended to filter out more examinees than RTE filtering. Rios et al. (2014), however, found slightly different results: they found small differences in estimates of performance between the two methods, and further that RTE filtering removed more examinee data than self-reported effort filtering.

Regardless of how examinee motivation is measured (e.g., self-report, RTE scores), practitioners must determine a cutoff criteria to use to discern the motivated from un- or low-motivated examinees. Sundre and Wise (2003) applied motivation filtering using SOS total scores, which range from 10 (least motivated) to 50 (most motivated), and filtered out examinee data associated with SOS scores as low as 20 and as high as 35. Other researchers have opted to use only the SOS effort subscale, which ranges from 5 (least effort) to 25 (most effort), and have filtered examinees with effort scores less than or equal to 10 (e.g., Wise & Kong, 2005), 13 (e.g., Liu, Rios, & Borden, 2015; Rios et al., 2014; Wise & Kong, 2005), or 15 (e.g., Swerdzewski et al., 2011). When filtering using RTE scores, researchers have generally used a cutoff of .90, meaning that examinees must have exhibited solution behavior on at least 90% of the items (e.g., Liu et al., 2015;
Rios et al., 2014; Swerdzewski et al., 2011; Wise & DeMars, 2010; Wise & Kong, 2005), although DeMars (2007) used a cutoff of .84.

Since its introduction in 2003, motivation filtering has become one of the most commonly used methods for addressing low motivation in low-stakes testing data. Studies using motivation filtering tend to support Sundre and Wise's (2003) initial findings that motivation filtering generally increases both group proficiency estimates and correlations with external variables, compared to results based on unfiltered data (for a review, see Table 1 of Steedle, 2014). Aside from motivation filtering, other post-hoc statistical techniques for addressing low examinee motivation include rapid-response filtering (Wise, 2006) and the effort-moderated IRT model (Wise & DeMars, 2006).

**Rapid-response filtering.** Rapid-response filtering (Wise, 2006) is similar to motivation filtering using RTE scores in that it relies on distinguishing solution behavior from rapid-response behavior at the item level. Unlike motivation filtering, however, rapid-response filtering removes data at the item level, rather than at the examinee level. Thus, rapid-response filtering is most appropriate for item-level analyses, such as examinations of item difficulty, where it provides the advantage of retaining some item-level data that may have otherwise been filtered out using examinee-level motivation filtering (e.g., if an examinee only exhibited solution behavior on a few items).

**Effort-moderated IRT model.** Wise and DeMars (2006) provide another approach for handling rapid responding; specifically, for handling rapid responding when the goal is to apply an IRT model to the data. The effort-moderated IRT model is similar to rapid-response filtering in that it, too, adjusts for item-level responses associated with rapid-guessing behavior. Essentially, this model acts as a three parameter logistic (3PL) model
for trustworthy data (i.e., for data associated with solution behavior), while acting as if the untrustworthy, rapid responses were unobserved. This model can be implemented using conventional IRT software (e.g., BILOG-MG) by re-coding rapid responses as “not administered” and performing item calibration with a 3PL model as usual (Wise & DeMars, 2006).¹ For data containing rapid responses, the effort-modernated IRT model has been shown to outperform the traditional 3PL with regard to model fit, accuracy of parameter estimates, accuracy of test information, and convergent validity of ability estimates (Wise & DeMars, 2006).

Although motivation filtering, rapid-response filtering, and the effort-modernated IRT model function in different ways, a common theme across these three post-hoc methods is that they deem data from low-motivated examinees untrustworthy. Further, these methods elect to work solely on the basis of the trustworthy data. They function as if the data associated with low-motivation were not even collected at all: as if they were missing.

**Missing Data**

As mentioned previously, untrustworthy data stemming from low-motivated examinees can be viewed as a missing data problem. When the data examinees provide are not reflective of their true proficiency, the data we want to analyze in order to make inferences about the examinee’s proficiency are effectively missing. In other words, although we may have observed data, these observed data are not what we are interested in., they are essentially decoys. The data we are interested in are missing. Due to its relevance to addressing low-motivated examinee data, the following section is dedicated

---

¹ This coding scheme approach was also used by Schnipke (1996) to handle rapid guessing on speeded tests.
to the topic of missing data. This section begins with a brief overview of missing data theory, follows with a description of the three missingness mechanisms, and concludes with a discussion of methods for handling missing data.

**Missing data theory.** Missing data theory helps researchers conceptualize and categorize the missingness they may encounter in their data by offering a framework for how missing data are connected to observed data. Missing data theory, which relies heavily on Rubin’s (1976) seminal work, considers three types of variables, and importantly, the relationships between them. The first type of variable is our variable of interest, typically referred to as ‘Y’. In practice, due to missing data, we will only observe a subset of Y, which I will refer to as ‘Y_{obs}’. The second is a variable indicating the missingness itself, typically referred to as ‘R’. R is a dichotomous variable governed by an underlying probability distribution. If a case is missing its Y value, R will traditionally have a value of one; if Y is not missing, R will traditionally have a value of zero. For example, Table 2 contains a variable of interest, Y_{obs}, with missing values, and a variable indicating missingness, R. Note that cases with missing Y values are denoted with ones for their R values. The third important variable is typically referred to as ‘X’, and represents a second measured variable (in addition to Y) that may or may not be of interest. The relationships between missingness (R) and the other variables (X, Y) are what distinguish the three missingness mechanisms. Although in practice we can never definitively determine which of these mechanisms are present, they serve as a useful framework for conceptualizing missingness and helping us decide how to best handle missing data.
**Missingness mechanisms.** Missing data mechanisms serve to describe missingness; specifically, what relates to the probability of missingness for a given variable. Three mechanisms are almost unanimously referred to in the missing data literature, namely, missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR). Each mechanism is described below.

Of the three mechanisms of missing data, data that are missing not at random (MNAR) are the most concerning. The MNAR mechanism is present when missingness, R, is related to the variable with missing data, Y, once controlling for X. R can also be related, or not, to X, and X can be related, or not, to Y; the important relationship for MNAR is that between R and Y, once controlling for X. The leftmost panel of Figure 3 displays the MNAR mechanism.

Data would be MNAR if, for example, examinees were to leave an item blank because they did not know the correct answer. In this scenario, missingness would be related to the value of the variable with missing data. We would expect that there would be a negative relationship between scores on that item (including those that were unobserved) and missingness on that item. That is, we would expect that examinees with missing values would have lower item scores (had they been observed) than examinees without missing values. This scenario may present in situations where examinees skip items that they are unsure of as a test-taking strategy if getting an item wrong counts negatively towards their score (and leaving the item blank neither helps nor hinders their score).

Data that are MNAR present a significant challenge to researchers because, as discussed later in this section, the most commonly used methods for handling missing
data assume that data are not MNAR. Such methods produce biased estimates when used with data that are MNAR. Of course, in practice we cannot definitively identify whether or not data are MNAR because such an identification relies on the counterfactual knowledge of the values of the missing data.

Data are missing at random (MAR) when the missingness, R, is not related to the variable with missing data, Y, once controlling for X, but is related to another measured variable, X (as is true for MNAR, X can be related, or not, to Y). The middle panel of Figure 3 displays the MAR mechanism. Consider the following example, which would produce data that are MAR: on a low-stakes mathematics exam, one item relies on information that students must glean from a reading passage about Orioles baseball players’ performance statistics. Because the reading passage is somewhat long, only students who are avid baseball fans tend to respond (uninterested students tend to skip the item). The data for this item would thus be MAR: missingness would not be related to the value of the variable with missing data, but it would be related to another measured variable (assuming, importantly, that baseball interest had been measured). We would expect that there would not be a relationship between scores on the mathematics item (including those that were unobserved) and missingness on that item, controlling for baseball interest. We would expect that examinees with missing values (i.e., students who are not avid fans) would not have different item scores (had they been observed) than examinees without missing values (i.e., student who are avid fans), controlling for baseball interest. Said differently, missingness would not be related to mathematics proficiency (controlling for baseball interest), but it would be related to interest in baseball.
Although, like the MNAR mechanism, we cannot definitively test for the presence of the MAR mechanism, we can test for relationships between measured variables and missingness. Consider our example; we could test for group mean differences in baseball interest between those who did and did not have missing data on the mathematics item (assuming we had measured baseball interest). For this example, we would expect those with missing data on the mathematics item to have a lower mean level of baseball interest than those without missing data on the mathematics item. Importantly, though, this test does not rule out the presence of the MAR or MNAR mechanism. It is possible, for example, that missingness on the mathematics item could also be related to the scores on the mathematics item (including those unobserved), even after controlling for baseball interest. A definitive test for MAR (or MNAR) requires knowledge of the relationship between the complete set of scores on the item (which we do not have) and the missingness on the item.

Finally, data are considered to be missing completely at random (MCAR) when the missingness, R, is neither related to the variable with missingness, Y, once controlling for X, nor any other measured variable, X (again, X can be related, or not, to Y). The rightmost panel of Figure 3 displays the MCAR mechanism. Data would be MCAR, for example, if different forms of a test (containing different items) were administered to a group of examinees and examinees were randomly assigned to receive a given test form. In this scenario, missingness would not be related to (a) the values of a given item with missing data, or (b) another measured variable. We would not expect the group means of examinees with missing data on a given item (i.e., examinees who were randomly selected to receive a form without that item) to differ from those of examinees who did
not have missing data (i.e., examinees who were randomly selected to receive a form with that item). Like the other two missingness mechanisms, we cannot definitively test for the MCAR mechanism. However, we can rule it out by testing for relationships between R and other measured variables (as described previously); if any of these tests result in significant relationships, then we can safely rule out the presence of the MCAR mechanism.

Even though in practice we will never know with certainty which missingness mechanism is at play for any given dataset with missing data, researchers should still consider which mechanism(s) may be more or less likely. Such considerations can inform the selection of additional variables to measure, often referred to as auxiliary variables.

**Auxiliary variables.** Auxiliary variables are used in conjunction with modern missing data techniques for the primary purpose of increasing the possibility that the data will be MAR (rather than MNAR). That is, we use auxiliary variables to try to account for a potential relationship between missingness and the variable of interest. Ideally, once controlling for the added auxiliary variables, any potential relationship between missingness and the variable of interest is eliminated. The best auxiliary variables are those that relate to either the missingness variable, or the full (unobserved) variable of interest with missing data, or both (Enders, 2010). Because the best performing missing data methods assume the data are MAR (as discussed in the next section), it is generally recommended that researchers incorporate any and all possibly relevant auxiliary variables in effort to stave off the MNAR mechanism. This liberal incorporation of auxiliary variables is referred to as an inclusive strategy (e.g., Collins, Schafer, & Kam, 2001).
Not all researchers agree with such a liberal use of auxiliary variables, though. For example, Hardt, Herke, and Leonhart (2012) warn that too many auxiliary variables can lead to biased estimates, and make a preliminary recommendation to have a minimum ratio of three complete cases per auxiliary variable. Additionally, Thoemmes and Rose (2014) caution that in some (perhaps rare) instances auxiliary variables can actually induce MNAR where MAR previously held. Nonetheless, generally speaking, the primary purpose of auxiliary variables is to attempt to upgrade a missingness mechanism from MNAR to MAR, which opens the door for the appropriate use of modern missing data methods. Thus, considering the missingness mechanism is important in determining the best method for handling missing data.

**Methods for handling missing data.** The problem of missing data is commonly addressed using one of three types of methods: deletion, single imputation, or one of the two ‘modern’ missing data techniques. Each of these methods is accompanied by certain assumptions. The former two methods tend to be easier to implement, yet require stricter assumptions about the missing data mechanism and even when those assumptions are met, they do not perform as well as the modern methods. The modern missing data methods are less restrictive in their assumptions about the missing data mechanism, yet are slightly more difficult to conceptualize and implement. The discrepancy in ease of implementation between the modern and less modern methods, however, has been greatly reduced since the inception of the more complex methods: many commonly-used software programs now include options to use the modern methods. In the sections that follow, each of the three types of missing data methods are described, with the most attention given to the last method, multiple imputation, as it is the focus of this study.
**Deletion.** Perhaps the easiest method for handling missing data is to discount, or delete, cases (e.g., examinees) associated with missing data. This can be accomplished using either pairwise or listwise deletion. Pairwise deletion removes cases on an analysis-by-analysis basis: only cases with complete data for the variables involved in a given analysis are analyzed. Thus, the sample of cases used can vary depending on the variables included in the analyses. Consider the example dataset in Table 3. If we were to conduct an analysis using variables Y₁ and Y₂ (e.g., a correlation), pairwise deletion would temporarily retain and perform analyses on cases 3, 4, 5, and 7 (sample size = 4). If we were then interested in variables Y₁ and Y₃, we would temporarily retain and perform analyses on cases 1, 3, 4, 5, and 7 (sample size = 5). One of the advantages of pairwise deletion is that it maximizes the sample size for a given analysis (i.e., you are not penalized for missing data on variables not of interest to the immediate analysis). One of the disadvantages of pairwise deletion is that it can lead to estimation issues (e.g., non-positive definite correlation matrices). Another disadvantage of pairwise deletion is that it assumes the data are MCAR. If this assumption is not met, as is often the case in practice, pairwise deletion can produce inaccurate parameter estimates.

Listwise deletion removes cases that have any missing data on any variable in the analysis.² Thus, if all correlations between the variables in Table 3 were desired, an analysis based on listwise deletion would use only cases with complete data on all variables (i.e., we would retain cases 3, 4, 5, and 7; sample size = 4). The primary advantages of listwise deletion are that it is easy to implement, and that it results in a

---

² Some researchers apply listwise deletion to an entire dataset, regardless of their intended analyses, so that only complete data on every variable in the dataset remain. This form of listwise deletion often results in the most drastic reductions in sample size.
complete dataset for a given analysis. The primary disadvantages are that the sample size of the dataset for any given analysis will often be dramatically reduced, and that, like pairwise deletion, listwise deletion assumes the data are MCAR. If this assumption is not met, any resulting parameter estimates may be inaccurate.

**Single imputation.** Rather than removing cases due to missing values, single imputation methods serve to replace missing values with a single plausible alternative value. It should thus follow that all cases are retained when using imputation methods. Different imputation methods take different approaches to determine the value to impute for any given missing data point.

Mean imputation replaces missing values for a variable with the average of the non-missing values for that variable.³ For example, consider variable \( Y_1 \) in Table 3. Mean imputation would impute a score of 7 for both case 2 and case 6, because the average of the observed data for \( Y_1 \) is 7 (i.e., \( \frac{12+4+8+2+9}{5} = 7 \)). Although it may be easy to implement, mean imputation can produce inaccurate parameter estimates and standard errors even when data are MCAR (e.g., artificially deflated variances due to imputing the same mean value repeatedly across an item), and has been deemed “possibly the worst missing data handling method available” (Enders, 2010, p. 43).

Rather than imputing the same value for all cases with missing data on a given variable (like mean imputation), regression imputation predicts the values of the missing data. These predicted values are then imputed into the dataset, resulting in a complete dataset. The imputed values are the result of a regression equation in which scores on the

---

³ It may be helpful for some readers to refer to this type of mean imputation as item-mean imputation, as to distinguish it from person-mean imputation. Person-mean imputation is a similar method wherein each person’s mean across all variables with non-missing values is imputed for each of his or her missing values. Neither type of mean imputation is recommended for use.
variable with missing data are predicted from other available (measured) complete variables. Because different cases may have different available complete data, depending on the pattern of missing data, multiple multiple regression equations are used: one for every missing data pattern in the dataset. For example, if we were to apply regression imputation to Y1 in Table 3, we would need to use two different equations: one predicting Y1 from both Y2 and Y3 (for case 6), and one predicting Y1 from Y2 only (for case 2). As a result of using linear regression, the imputed values for a given variable will all fall directly on the regression line, indicating perfect prediction. In reality, perfect prediction would not occur. Unfortunately, even under the MCAR mechanism, regression imputation tends to produce biased (underestimated) variance estimates and standard errors for the variable with missing data, and thus is not recommended (Enders, 2010).

Stochastic regression imputation improves upon regression imputation by supplementing the regression equation(s) with a stochastic (i.e., random) error term when predicting values of the variable with missingness. The role of the random error term, which is normally distributed with a mean of zero and a variance equal to the residual variance from the regression imputation model, is to infuse variability into the imputed values. When data are MAR, stochastic regression imputation produces unbiased parameter estimates. Because of this, stochastic regression imputation is perhaps the best choice out of the traditional missing data methods. However, it is not without flaw; like the other single imputation methods, stochastic regression imputation still produces biased (underestimated) standard errors.4

---

4 If this seems counter-intuitive, consider that there are two important types of variability: variability within a sample, and variability between samples. The stochastic term in the regression equation addresses the former, while the latter is what leads to underestimated standard errors.
Modern missing data methods. Two techniques, full information maximum likelihood estimation (FIML) and multiple imputation (MI), are often grouped together and referred to as the ‘modern’ missing data methods. Unlike the traditional deletion or single imputation methods, FIML and MI are slightly more complex with regard to how they work, and are slightly more difficult to implement. Nonetheless, these methods are incredibly useful because they produce both unbiased parameter estimates and unbiased standard errors when data are MAR. Additionally, the two methods yield asymptotically equivalent results (Collins et al., 2001).

Unlike all of the other missing data methods discussed here, full information maximum likelihood (FIML) neither deletes nor imputes data. FIML uses all of the available data to answer the question, “What parameter values are most likely to have produced the observed data?” That is, FIML evaluates the likelihood of the observed data (with missingness) across different parameter values and determines which parameter values result in the greatest likelihood of having produced the observed data. Note that FIML does not require any pre-processing of the missing data; the analyses of interest are performed directly on the observed dataset, missingness and all. The primary advantages of FIML are that it can be fairly easy to implement (many software programs include FIML as an estimator), and like multiple imputation, FIML produces unbiased parameter estimates and standard errors when data are MAR. Some of the disadvantages of FIML are that it can become increasingly complex and difficult to implement when (a) auxiliary
Multiple imputation (MI) is a missing data method in which multiple, plausible values are imputed for each missing data point for a given variable in a dataset. Like single imputation methods, MI results in complete data. Unlike single imputation methods, MI results in *multiple* complete datasets, with each complete dataset differing only in the imputed values. Conceptually, you can think of MI as stochastic regression imputation performed multiple times (resulting in multiple, complete datasets). MI is performed in two phases: first, an imputation phase, and second, an analysis and pooling phase. Each of these phases are discussed, in turn, below.

The imputation phase of MI begins with the single observed dataset (with missing data), and ends with multiple versions of complete, imputed datasets. Although the imputation phase of MI can employ various algorithms, the data augmentation (DA) algorithm is perhaps the most commonly used and the most commonly available in software packages (Enders, 2010). Like FIML, the DA algorithm assumes the data are MAR and follow a multivariate normal distribution (Enders, 2010). The DA algorithm consists of two steps, an imputation step (I-step) and a posterior step (P-step), with an imputed, complete dataset resulting from multiple iterations of the two steps. Essentially, the I-step uses stochastic regression imputation to impute plausible values for the missing data, and then the P-step uses the imputed, complete data from the I-step to re-estimate the mean and variance of the data. The resulting, updated mean and variance estimates

---

* It is important to note that these disadvantages are with regard to the use of FIML through the traditional structural equation modeling (SEM) framework, and may not apply to FIML used through an IRT framework.

* Although some authors present analysis and pooling as two distinct phases (e.g., Enders, 2010; in press).
are then fed back to the I-step, the regression equations are updated (based on the new mean and variance estimates), and the missing data are re-imputed. The DA algorithm performs I- and P-steps iteratively, ultimately creating a chain of imputed datasets.

Again, this is essentially how the DA algorithm works to complete the imputation phase of MI; but of course, in practice the I- and P-steps are a bit more complex.

The first I-step begins with an initial set of estimates for both a vector of means and a covariance matrix (recall that we are assuming our data follow a multivariate normal distribution). FIML is commonly used to generate these initial estimates (i.e., starting values). The initial mean vector and covariance matrix are then used to derive the parameter estimates (e.g., regression coefficients and error variances) for multiple stochastic regression equations, one for each missing data pattern, where the residual error term is normally distributed with a mean vector of zero and covariance matrix equal to the residual covariance matrix from the regression of the variable with missing values on the other available variables in the dataset. The stochastic regression equations are then used to predict and impute a set of plausible values for the missing data. This imputed, complete dataset provides the basis for the first P-step.

The first P-step uses the imputed, complete data from the first I-step to re-estimate a vector of means and covariance matrix for the data. These parameter estimates are then used to generate posterior distributions for the parameters (i.e., a posterior for the means and a posterior for the covariances). New estimates of the parameters are then sampled randomly from each posterior. For the covariance matrix, this means that a new covariance matrix is randomly sampled from the posterior distribution of the covariance matrix which follows an inverse Wishart distribution governed by $N-1$ degrees of
freedom and the sums of squares and cross products matrix from the imputed dataset (where \( N \) is the sample size of the dataset). Similarly, a new vector of means is randomly sampled from the posterior distribution of the mean vector which follows a multivariate normal distribution with a mean vector equal to the vector of means that was estimated based on the imputed data, and a covariance matrix equal to the covariance matrix that was randomly sampled just prior. Generally, the purpose of the P-step is to provide new estimates of the parameters in the mean vector and covariance matrix, where the new estimates can be thought of as random perturbations from the previous estimates.

Following the first P-step, a second I-step begins, but this time the parameter estimates from the previous P-step are used to derive the coefficients for the stochastic regression equations. The I-step then re-implutes the missing values, resulting in a second complete dataset which is then fed to the second P-step.\(^7\) The I- and P-steps are performed iteratively, resulting in a chain of imputed datasets and parameter estimates that are each dependent on those immediately prior. Researchers generally only retain a fraction of the total number of imputed datasets from a long chain of I- and P-steps, however. Imputed datasets in the chain are discarded for two primary reasons, either because: (1) they occur at the very beginning of the chain, or (2) they share a dependency with an imputed dataset that has been retained. A common method for selecting imputed datasets to retain is to save the result from every \( i^{th} \) I-step, where \( i \) is determined by the number of iterations of the chain needed to obtain datasets that are not autocorrelated (e.g., retain the imputed datasets from every 100\( ^{th} \) I-step in the chain). Although early

\(^7\) Note that the coefficients for the stochastic regression equation differ across I-steps. Thus although conceptually speaking MI is akin to performing stochastic regression multiple times (using the same regression coefficients), we see that in practice these two methods are not equivalent.
guidelines suggested that retaining a total of only three to five imputed datasets were sufficient for MI (e.g., Rubin, 1996), more recent research has supported the retention of a greater number of imputed datasets (e.g., 20-100), especially when the fraction of missing information is high (Graham, Olchowski, & Gilreath, 2007; Johnson & Young, 2011). A definitive guideline for the number of imputations necessary has yet to be established, though. Once the desired number of imputed datasets are created, the imputation phase of MI is complete.

The analysis and pooling phase of MI begins with the multiple, imputed datasets generated by the imputation phase, and ends with a single set of parameter estimates of interest. The analysis portion of this phase is quite straightforward: the substantive analyses of interest are performed on each imputed dataset separately. For a total of 20 independent imputed datasets, this would mean performing the analyses of interest 20 times (once per dataset). The results of each of these analyses must then be pooled together to generate a single set of results. As an example, let us say we were interested in the relationship between scores on a measure of critical thinking and SAT total scores. After performing the imputation phase of MI, we end up retaining a total of, say, 50 independent imputed datasets. We would then perform our analysis of interest (e.g., bivariate regression) on each of the 50 datasets. This would give us 50 sets of results: 50 sets of parameter estimates (e.g., intercepts, slopes), and 50 sets of standard errors. To pool the parameter estimates, we simply take the average (Rubin, 1987), as shown in the equation below,

\[ X_{pooled} = \frac{\sum_{i=1}^{m} X_i}{m} \]  

(1)
where \( m \) is the number of imputed datasets, and \( X_i \) is the parameter estimate resulting from the analyses of imputed dataset \( i \). Calculating a pooled standard error is a bit more involved.

To pool the standard errors, two pieces of information are combined: information regarding the within-imputation variance, and information regarding the between-imputation variance (Enders, 2010). The variance within imputations is calculated as the average of the squared standard errors across datasets (i.e., the average within-imputation parameter variance), as shown in the equation below,

\[
\sigma^2_{\text{within}} = \frac{\sum_{i=1}^{m} SE_i^2}{m} \tag{2}
\]

where \( m \) is the number of imputed datasets, and \( SE_i^2 \) is the squared standard error of imputed dataset \( i \). The variance between imputations is calculated as the overall variance among the parameter estimates from each imputed dataset, as shown in the equation below,

\[
\sigma^2_{\text{between}} = \frac{\sum_{i=1}^{m} (X_i - \overline{X})^2}{m - 1} \tag{3}
\]

where \( m \) is the number of imputed datasets, \( X_i \) is the parameter estimate resulting from the analyses of imputed dataset \( i \), and \( \overline{X} \) is the average parameter estimate across the \( m \) datasets (see Equation 1). For our example, the within-imputation variance would be the average of the squared standard errors from each of the 50 analyses, and the between-imputation variance would be the variance of the 50 regression coefficients. The within- and between-imputation variances are then combined to create the pooled standard errors. Specifically, the pooled standard error is equal to the square root of the
sum of three terms: the within-imputation variance, the between-imputation variance, and the between-imputation variance divided by the number of imputed datasets (Enders, 2010), as shown in the equation below,

\[
SE_{\text{pooled}} = \sqrt{\sigma^2_{\text{within}} + \sigma^2_{\text{between}} + \frac{\sigma^2_{\text{between}}}{m}}
\]  

(4)

where \( m \) is the number of imputed datasets, and the within and between variances are as defined in the previous equations.

Although FIML and MI have been shown to produce asymptotically equivalent results (Collins et al., 2001), practitioners may prefer to use MI over FIML. One reason for this may be because MI provides complete datasets, rendering subsequent analyses on each dataset no different than if there had been no missing data (although the results must then be pooled). Further, as stated previously, MI is often preferred over FIML (as applied through the SEM framework) when either (a) auxiliary variables are included (Eekhout et al., 2015), or (b) data are missing at the item level (Gottschall et al., 2012).

Throughout this discussion of missing data one assumption has continuously been made: that the missing data are missing because they were not supplied by the examinee (i.e., examinee nonresponse). That is, that the data came to us with missingness already in place. In the next section I begin by discussing instances where missingness is not organic, but instead it is imposed on the data. In doing so, I transition to the topic of reframing post-hoc methods for addressing low examinee motivation as missing data methods.
Reframing Post-hoc Approaches to Examinee Motivation as Missing Data Methods

Recall that one of the purposes of this dissertation is to reframe post-hoc approaches to examinee motivation as missing data techniques, and to consider the use of the modern missing data method of MI to address examinee motivation. Before illustrating the connection between post-hoc approaches to examinee motivation and missing data techniques in this section, it is important to make the distinction between organic missingness and imposed missingness. Organic missingness, or missingness that “comes with” the data, is the context in which the missing data techniques described in the previous section are typically applied. Imposed missingness is missingness created by the data analyst. When we use post-hoc approaches to examinee motivation, we are converting non-missing data to missing, thus we are imposing missingness. The distinction between organic and imposed missingness has important implications when considering the missingness mechanism and thus the appropriateness of various missing data techniques for addressing examinee motivation.

After distinguishing between organic and imposed missingness, I then illustrate the connection between the post-hoc approaches to examinee motivation described in the first section of this chapter (e.g., motivation filtering, effort moderated IRT, rapid response filtering) with the missing data techniques described in the second section. Previous research comparing the various approaches is then summarized followed by considerations in the use of auxiliary variables in this context with the modern approaches.

Organic vs. imposed missingness. Generally, data are missing from datasets because they were never collected in the first place (e.g., examinees failed to provide
responses to some or all items that were administered to them). In this situation, the missingness is organic. There are some instances, however, when data are missing because they have been removed from the dataset. In other words, missingness is imposed. Perhaps the most common example of this occurs when researchers delete outlying data. Although it is not always the preferred method for dealing with outliers, researchers often delete outlying data when such data are suspected of not belonging to the same population as other supplied data (e.g., if the data fall outside the range of possible or plausible values for the associated variable). Kim, Reiter, Wang, Cox, and Karr (2014) describe a method for identifying such “erroneous or inconsistent” values, deleting them (resulting in ‘blanked’ data), and using a form of MI to impute them (p. 375). They describe this as a process for editing faulty data.

The idea of data editing and data blanking can be applied to the context of examinee motivation. Similar to how Kim et al. (2014) found data that failed to meet logical standards untrustworthy (e.g., “an establishment reporting an entry year of 2012 that also reports nonzero employment in earlier years”, p. 375), we find data from unmotivated examinees untrustworthy. Accordingly, such data is suitable for removal from a dataset, resulting in what we refer to as ‘motivation-imposed missingness’. Motivation-imposed missing data can be handled using any of the missing data methods previously discussed (e.g., listwise deletion), although some methods may be more appropriate than others. Examples of both traditional and modern missing data methods as applied to motivation-imposed missingness are discussed next, and the connection between these missing data methods and various post-hoc approaches to examinee motivation are made explicit.
Handling motivation-imposed missingness with traditional missing data methods. At present, the most common methods for handling motivation-imposed missingness are pairwise and listwise deletion. Recall that pairwise deletion involves dropping cases with missing data on an analysis-by-analysis basis. An example of pairwise deletion is realized in rapid-response filtering. Rapid-response filtering first imposes missingness on examinee responses that are not associated with solution behavior (i.e., motivation-imposed missingness), then performs subsequent analyses with all the complete data that are available on an analysis-by-analysis basis. Although rapid-response filtering may preserve the sample size for some analyses (compared to listwise deletion, for example), recall that it relies on the strict assumption that the missing data are MCAR. Deviations from this assumption lead to biased results. However, before I discuss the plausibility of the MCAR mechanism in the context of motivation-imposed missingness, let us first discuss the use of listwise deletion for handling motivation-imposed missingness, which also makes this assumption.

Listwise deletion involves dropping cases that have missing data on any variable being analyzed. An example of listwise deletion is realized in motivation filtering. Unlike rapid-response filtering, which filters at the item level, motivation filtering works by filtering (i.e., removing data) at the examinee level. Motivation filtering imposes missingness at the case level using measures such as self-reported motivation scores (e.g., SOS scores) or response time scores (e.g., RTE). The entire response strings of examinees who do not demonstrate the minimum acceptable motivation for the measure or testing period in general are deleted—even if some, or in some cases the majority of, responses in an examinee’s response string are associated with solution-behavior! Like
pairwise deletion, listwise deletion assumes the data are MCAR. Thus, like rapid-
response filtering, motivation filtering assumes the data are MCAR.

Recall that in order to be considered MCAR, the missingness (R) cannot be
related to the full set of values on the variable of interest (Y), nor can it be related to any
other measured variable (X). In the context of motivation-imposed missingness,
missingness is whether or not examinee responses are filtered. So, in order to satisfy the
MCAR assumption, missingness cannot be related to the full set of values on the variable
of interest, meaning that whether or not a response is filtered cannot relate to the
complete values of that variable. Although we may be tempted to test for a relationship
between the missingness and the values of the variable prior to imposing missing, we
must remember that the values for which we are applying motivation-imposed
missingness have been deemed untrustworthy. Thus, the full set of trustworthy values for
the variable of interest are truly missing (whether or not we impose missingness). It may
then make more sense to consider the relationship between missingness (i.e., motivation)
and examinee ability measured using something other than the test in question. For
instance, the relationship between missingness (motivation) and SAT math scores might
be examined when the test in question is a math test. Unfortunately, there is research to
both support (Rios et al., 2017; Wise et al., 2009; Wise & Kong, 2005\(^8\)) and negate
(Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005) the claim that
motivation is related to examinee ability (see Steedle, 2014 for a review), and therefore
there is research to both support and negate the claim that missingness is related to the

\(^8\) Wise and Kong (2005) found a small positive correlation between self-reported effort scores and SAT verbal scores; self-reported effort scores did not relate to SAT quantitative scores.
full (unobserved) set of trustworthy values for the variable of interest in the context of motivation-imposed missingness.

In considering the second half of the MCAR assumption, we consider the relationship between missingness (i.e., motivation) and other measured variables. In situations where we are filtering data at the item level, the MCAR assumption may be violated when multiple items are used to measure the same construct. This may occur when the items assess the same ability and this ability is related to motivation. Further, even if motivation is unrelated to ability, it may be related to other variables in our dataset. As educational researchers, we commonly include variables such as gender and grade level in our substantive analyses as covariates. Variables such as these have been shown to relate to examinee motivation (e.g., Wise, Ma, Kingsbury, & Hauser 2010).

Ultimately, in the context of motivation-imposed missingness, I believe the assumption of MCAR is likely untenable. It is likely that missingness will relate to either other measured variables in the dataset, or the full (unobserved) trustworthy values of the variable of interest, or both. Therefore, methods that assume the data are MCAR are not appropriate. Fortunately, modern missing data methods do not make this strict assumption: instead, they assume the data are MAR.

**Handling motivation-imposed missingness with modern missing data methods.** Although they have seen limited use, both modern missing data techniques have been used in the literature to handle motivation-imposed missingness. With regard to FIML, two studies have applied a version of this technique to handle motivation-imposed missingness. The first study is Wise and DeMars’ (2006) introduction of the effort-moderated IRT model. Recall that this model treats rapid responses as if they came
from items that were not administered (i.e., as if they are missing), and uses maximum likelihood estimation to estimate the model parameters based solely on the remaining observed data. This is essentially equivalent to using FIML (Wise & DeMars, 2006). The second study to use a FIML-like method comes from Rios et al. (2017), who applied a modified version of the effort-moderated IRT model, although the authors did not label it as such. Thus, the modern missing data technique of FIML has been used as a post-hoc method to address examinee motivation in situations where item-level missingness has been created on the basis of examinee effort and an IRT model using FIML estimation is then applied. Although the use of FIML to address missing data is not limited to IRT modeling, at present that is the only application of the technique in the context of motivation-imposed missingness.

With regard to the second modern missing data method, MI, the only attempt at applying this technique to the context of examinee motivation comes from Koepfler et al. (2011). Importantly, unlike the two FIML studies, Koepfler et al. (2011) are explicit about their use of a modern missing data method for addressing low examinee motivation. Because both Rios et al.’s (2017) and Koepfler et al.’s (2011) studies compared the performance of modern missing data methods to traditional missing data methods, more detail is provided about these studies, and their findings, below.

**Research comparing the use of traditional and modern missing data methods for addressing motivation-imposed missingness.** The purpose of Rios et al.’s (2017) study was to compare the performance of two methods for handling rapid responses, namely, examinee-level filtering (i.e., listwise deletion) and response-level filtering (i.e., FIML). First, the authors simulated item response data with varying levels of rapid
responses to compare the accuracy of parameter recovery for both methods. In some conditions, data were simulated such that rapid responding was related to true ability; in other conditions, data were simulated such that rapid responding was not related to true ability. Once the data were simulated, they were filtered either at the examinee or response level. Examinee-level filtering was performed by listwise deleting any examinee with rapid responses and analyzing the resulting complete dataset. Response-level filtering was performed by recoding rapid responses as blank and using a modified 3PL model (i.e., a 3PL model with the lower asymptote parameter held constant at 0.25) to analyze the incomplete dataset. The results of the simulation study indicated that when rapid responding was not related to ability, both filtering methods performed equally as well in recovering the examinees’ average performance on the test. However, when rapid responding was related to ability, response-level filtering outperformed examinee-level filtering, with examinee-level filtering producing significantly inflated group mean estimates.

Following the results of the simulation study, Rios et al. (2017) conducted a similar study with real data. Again, they compared both examinee-level filtering (i.e., listwise deleting examinees with more than a given number of rapid responses and analyzing the subsequent complete dataset), and response-level filtering (i.e., deleting rapid responses and using the modified 3PL to analyze the incomplete dataset). For the applied dataset, examinee-level filtering resulted in estimates of average test performance that were higher than those from both the total, unfiltered data and those from the response-level filtered data. Because this pattern of results aligns with the results from the simulation study in which ability was related to effort, the authors suggest that these
results are indicative of a relationship between effort and ability. This suggestion was further supported by significant differences in SAT scores between the motivated and unmotivated examinees in the applied dataset.

Similar to Rios et al. (2017), Koepfler et al. (2011) used both simulated and real data to compare the performance of different missing data methods for handling low-motivated examinee responses. Unlike Rios et al. (2017), though, they used MI (rather than FIML) as their modern missing data method. For their simulation study, Koepfler et al. (2011) compared the performance of four methods in recovering item parameters and group mean performance when data were MCAR and MAR, namely, (1) scoring rapid responses as incorrect responses, (2) listwise deleting simulees with more than a given number of rapid responses, (3) treating items with rapid responses as if they had not been administered to the simulee, and (4) converting rapid responses to missing (i.e., motivation-imposed missingness) and applying MI to the missing responses. The authors found that MI recovered the group mean performance parameter as well as, or better than, the other three methods, and that MI recovered item parameters (i.e., difficulty and discrimination) with little to no bias.

For their study with real data, the authors began by deleting any item responses associated with rapid responding. They then compared the resulting item parameters and group mean estimates from four methods, namely, (1) listwise deleting examinees with any missing responses (even those with only a single missing response), (2) listwise deleting examinees with 10% or more missing responses (i.e., motivation filtering), (3) retaining the original, rapid responses, and (4) applying MI to the missing responses. Generally, the authors found that the more extreme listwise deletion method (i.e., method
one) resulted in the highest item difficulty and group mean estimates, followed by the less extreme listwise deletion method (i.e., method two), followed by MI. Retaining the original responses resulted in the lowest item difficulty and group mean estimates, supporting the research that indicates a relationship between test performance and motivation. Koepfler et al. (2011) ultimately concluded that the method used to handle missing/rapid responses impacts the inferences made from the data, and further that of the methods examined, MI performed the best.

Together the results of these studies, which both compared the performance of traditional and modern missing data methods in addressing low examinee effort, suggest similar things. First, both studies suggest that analyzing data that include rapid responses (without altering the data at all) tends to produce underestimates of examinee group performance, and that listwise deleting or motivation filtering data that include rapid responses tends to produce overestimates of examinee group performance. These tendencies were most pronounced when the data were MAR, and as more data were missing. Thus, neither of these methods were recommended for use in this context. Second, the results of the applied portions of both studies were suggestive of a relationship between rapid responding and ability. That is, the results of the real data applications most aligned with the simulation results in which the data were simulated to be MAR with rapid responding related to ability. Ultimately, after comparing traditional and modern methods for handling rapid responses using both real and simulated data, both sets of authors advocated for the use of modern methods over traditional methods.

**Auxiliary variables in the context of motivation-imposed missingness.** The results of the Rios et al. (2017) and Koepfler et al. (2011) both indicate a strong
preference for the use of modern missing data techniques for addressing low examinee motivation. Because auxiliary variables can be used with modern missing data techniques to make the MAR assumption more plausible, it is important to consider what types of auxiliary variables are best suited for use in this context. Recall that the best auxiliary variables are those that relate to either the missingness variable, or the full (unobserved) variable of interest with missing data, or both (Enders, 2010). In the context of motivation-imposed missingness, this implies that auxiliary variables should either relate to examinee motivation (i.e., missingness), or to examinee ability (i.e., the variable of interest), or both. Although it may seem appropriate to include the measure of examinee motivation used to impose missingness into the model (e.g., item response time), we argue that in this context it is unwise. Consider the fact that auxiliary variables are used to predict the missing values (i.e., they are included in the regression equations of MI). If we only consider the relationship between item response time and ability for responses associated with solution behavior, we might see a negative relationship. That is, given that examinees exhibited solution behavior for an item, longer response times may be associated with lower abilities. Thus, if we extrapolate this relationship to the missing data points that are associated with short response times, we might be predicting very high abilities for the rapid responders (as noted by Wise & DeMars, 2005). We must keep in mind that we want to predict examinee responses *as if the examinee had demonstrated solution behavior*; thus we do not want to predict item responses based on the observed response times that are associated with the untrustworthy, removed responses. The observed response times associated with the untrustworthy, rapid responses are not
indicative of the time it would take examinees to fully consider the items and select appropriate responses.

Given that it is unwise to include auxiliary variables that relate to motivation, in this context, we should aim to include auxiliary variables that relate to examinee ability pertaining to the measure of interest. Recall that in the examinee motivation literature, researchers often operationalize ability by SAT scores. These researchers are likely assuming that SAT performance is related to the ability being measured by the measure of interest. Therefore, SAT scores may be useful as auxiliary variables in the context of motivation-imposed missingness.

Some of the best auxiliary variables for motivation-imposed missing data, however, are likely the responses to items for which examinees exhibit solution behavior. Because we have evidence that examinees tend to display varying levels of motivation while completing a measure (Wise & Kong, 2005), we can use the responses for which examinees exhibited solution behavior to predict and impute responses that were previously associated with rapid responding. Further, if all of the items are measuring the same ability, we are essentially controlling for ability by including them as auxiliary variables. Thus, in the event that examinee motivation is related to ability (i.e., that the data are MNAR), we can essentially control for this by treating the other items as auxiliary variables—effectively shifting the data to be MAR. Said differently, there should not be a relationship between motivation and the responses that would have been observed had all of the examinees exhibited solution behavior (i.e., between R and Y) after accounting for the relationship between motivation and the observed solution-behavior responses (i.e., between R and X).
By including auxiliary variables into MI, the chance of the data being MAR rather than MNAR increases, and thus the appropriateness of MI (or FIML) over traditional missing data methods increases. Generally, the purpose of this study is to compare the application of MI to data that have motivation-imposed missingness to other more commonly used post-hoc methods for addressing examinee motivation.

Purpose of Study

Recall that this dissertation has two primary purposes, one didactic in nature and one empirical in nature. The first, reframing the problem of examinee effort as a problem of missing data, has been addressed throughout this chapter. The second purpose of this dissertation is addressed through an empirical study.

The purpose of the empirical study is to compare three post-hoc methods for addressing low examinee motivation using real data. The dataset I analyzed contains both solution-behavior responses and rapid responses. The three post-hoc methods that were applied are (1) leaving the data as they were observed, complete with a mix of rapid and solution-behavior responses, (2) motivation filtering (i.e., listwise deleting examinees with more than an acceptable amount of rapid responses), and (3) using MI with auxiliary variables to impute plausible solution-behavior responses in place of any and all rapid responses. The first method is often seen in practice in situations where researchers do not consider the motivation of the examinees. The second, motivation filtering, is one of the most commonly used post-hoc approaches for handling data collected in low-stakes testing contexts. The third method, MI, is one of the modern missing data methods, which have seen little use as applied to datasets with rapid responses.
I note two studies that have applied modern missing data techniques for the treatment of rapid responding: Koepfler et al. (2011), in which MI was applied, and Rios et al. (2017), in which a method similar to FIML was applied. This study expands Koepfler et al.’s (2011) work by examining the impact of missing data method on relationships between growth estimates and other variables of interest, and by including auxiliary variables into the MI procedure. This study expands Rios et al.’s (2017) work by reframing the issue of rapid responding as a missing data problem, and by using MI, rather than an IRT-based approach, to handle the rapid responses. This study expands on both Koepfler et al.’s (2011) and Rios et al.’s (2017) work by considering the performance of missing data methods in estimating growth, rather than estimating ability at a single time point.

Research Questions

The analyses performed for this study, which are described in the next chapter, are intended to address four research questions.

Research question one. The first research question addresses the plausibility of a relationship between examinee effort and ability, and examinee effort and other measured variables, for our data. Traditional post-hoc methods like motivation filtering assume examinee effort is neither related to (a) examinee ability, nor (b) any other measured variable in the dataset (i.e., that the data are MCAR), yet there is no way to directly test this assumption because it relies on counterfactual data. However, we can indirectly examine this assumption in this study by comparing (a) item-level solution-behavior responses, and (b) other measured variables (e.g., SAT scores), for examinees categorized as demonstrating adequate test-taking effort and examinees categorized as not
demonstrating adequate test-taking effort. Formally, research question one asks: How do solution-behavior item scores and other auxiliary variable scores from examinees demonstrating adequate test-taking effort (i.e., those with RTE scores $\geq .90$) compare to those from examinees demonstrating low test-taking effort (i.e., those with RTE scores $< .90$)?

**Research question two.** The second research question is intended to address the uncertainty regarding the recommended number of imputations within MI. Because there is little research investigating the impact of the number of imputations on the results of MI (e.g., Graham et al., 2007; Johnson & Young, 2011), and because what little research exists does not agree on a specific recommendation, for this study, the results of MI were compared across different numbers of imputations. Due to the applied nature of this study, we examined the results with regard to the pooled standard errors. Specifically, research question two asks: How do the pooled standard errors compare across different numbers of imputations for MI?

**Research question three.** The third research question addressed by this study has to do with the impact of post-hoc method on inferences made regarding student growth over time. Formally, research question three asks: How do the statistical significance conclusions regarding the effect of time on NW-9 scores compare across the three post-hoc methods for handling examinee motivation (i.e., leaving the data intact, motivation filtering, and using multiple imputation to impute plausible replacement values for rapid responses)?

**Research question four.** Like the third research question, the fourth research question also addresses the impact of post-hoc method on inferences made regarding
student growth over time. This research question builds on the previous research question, though, by examining differences across methods with regard to a suspected predictor of growth in student learning. Specifically, research question four asks: How do the statistical significance conclusions regarding the interaction between time and the number of relevant courses completed at post-test (a suspected moderator of growth in student learning) on NW-9 scores compare across the three post-hoc methods for handling examinee motivation (i.e., leaving the data intact, motivation filtering, and using multiple imputation to impute plausible replacement values for rapid responses)?
CHAPTER THREE

Method

Introduction

The purpose of this chapter is to describe the methods that were used to conduct this study. This chapter begins with a discussion of how and from whom data were collected for this study, as well as a description of the measure of interest (i.e., the NW-9). I then describe the use of item response times as measures of examinee effort, and the process used to distinguish solution-behavior responses from rapid responses (i.e., the CUMP method). Finally, the specific analyses that were performed to address each of the four research questions are described.

Data Collection

The data analyzed for this study were collected from undergraduate students at James Madison University (JMU) at two time points during the students’ tenure at JMU. The first time point occurred during the students’ first-year orientation week, prior to completing any coursework at JMU (i.e., at the start of the Fall semester of their first year at JMU). The second time point occurred after the students had completed between 45 and 70 credit hours at JMU (during the Spring semester of their sophomore year). This pre-post design is used to help determine students’ knowledge, attitudes, and abilities both before and after completing aspects of JMU’s General Education program.

For both Fall and Spring time points, the data were collected as part of JMU’s Assessment Days. On Assessment Days, eligible students are required to come to assigned testing rooms to complete a series of assessments during a 2- to 3-hour standardized, proctored testing session. Not all students complete the same assessments
during a given Assessment Day. During the Fall Assessment Days, different test configurations are assigned to different testing rooms based on the sample size and administrative requirements for each test. Students are randomly assigned to testing rooms, however, and thus are randomly assigned to tests as well. During the Spring Assessment Days, test configurations and students are assigned to the same rooms as they were for the associated Fall Assessment Day (e.g., the Spring 2017 assignments reflect the Fall 2015 assignments). This helps to ensure that students complete the same assessments at both time points.

Assessment Day testing at JMU is low-stakes for students, meaning that students are not directly impacted by their performance on the assessments. In an effort to motivate students to complete the assessments to the best of their ability, the purpose and importance of the assessments are explained to students prior to and on the day of Assessment Day. Further, the testing sessions are conducted by proctors who are aware of the importance of the assessment data, and are trained to encourage students to attend to the assessments.

**Examinees**

The dataset analyzed for this study contains complete pre- and post-test data from 388 students, across three cohorts. Specifically, 159 students completed the assessment in the Fall of 2008 and again in the Spring of 2010, 99 students completed the assessment in Fall 2010 and again in Spring 2012, and 130 students completed the assessment in Fall 2011 and again in Spring 2013. All 388 completed the measure of interest electronically, which allowed for the associated item-level response times to be recorded. Each student

---

9 NW-9 scores were compared by cohort prior to combining the data; there were neither statistical nor practical differences in scores across the three cohorts.
in this sample has complete item-level response time information at both pre- and post-test.

With regard to demographic makeup, the sample of students used for this study is representative of the population of students at JMU. For sample descriptive information, see Table 4.

Measures

Three types of measures were used for this study: (1) the measure of substantive interest, the Natural World Test version 9 (NW-9; Sundre, 2008), (2) the measure of examinee effort, students’ item-level response times, and (3) a set of measures that served as auxiliary variables.

The Natural World Test version 9. The NW-9 is a 66-item measure of undergraduate students’ scientific and quantitative reasoning ability. It was developed by faculty at JMU for the purpose of providing “information about the effects of curriculum and instruction on students’ learning” (Sundre, 2008, p. 3). The NW-9 items are used to assess, and are mapped to, eight student learning objectives within Cluster 3 of JMU’s General Education program. The item-to-objective map has been supported by backwards translation and content alignment activities performed by content experts (Sundre, 2008).

All 66 items are multiple choice, with between two and six response options per item. Students are given 60 minutes to complete all 66 items. All but one item have one correct response option; item 24 has two correct response options. All items are scored as correct (coded ‘1’) or incorrect (coded ‘0’). Higher total scores indicate higher levels of scientific and quantitative reasoning ability.
The internal consistency reliability (alpha) of NW-9 scores was estimated to be .73 for first-year students in the Fall of 2007, and .65 for sophomores in the Spring of 2008 (Sundre, 2008).\textsuperscript{10} For the observed responses analyzed in this study, the internal consistency reliability (alpha) estimates were .74 for the Fall sample, and .80 for the Spring sample. Various forms of validity evidence have been collected for the NW-9, including the backward translation and content alignment work described by Sundre (2008), and Sundre and Thelk (2010). Sundre and Thelk (2010) reported a positive relationship between the number of related General Education courses JMU students had completed and their NW-9 scores, as well as a positive relationship between JMU students’ grades in Cluster 3 courses and NW-9 scores. Both Sundre and Thelk (2010) and Hathcoat, Sundre, and Johnston (2015) reported increases in students’ NW-9 scores from students’ first year to their sophomore/junior year.

**Examinee effort.** Examinee effort was measured at each time point by examinees’ item-level response times. This information was used to inform both the motivation filtering and multiple imputation procedures (item response times will not be considered for the method in which the data are left intact). Item response times were recorded by the test administration software (Adaptex; Wise & Yang, 2003), and indicate the time that elapsed from when an examinee is first presented an item to when they complete the item.

For each item at each time point, a threshold was set in order to distinguish between responses provided using solution behavior and rapid responses. Although multiple methods for determining thresholds have been studied and supported (Kong et

\textsuperscript{10} These reliability estimates represent scores collected via computer-based administration only (because only scores from computer-based administration will be used for this study).
al., 2007), for this study item response time thresholds will be determined using the cumulative proportion (CUMP) method (Guo et al., 2016), as described previously. As was done by Guo et al. (2016), the chance correct function was the inverse of the number of response options for a given item (e.g., for an item with 4 response options, there is an associated 1/4, or .25, probability of responding correctly by chance). Because the cumulative proportion correct function may intersect with the chance correct function more than once (e.g., due to small sample sizes at the lower response time values), we selected the intersection associated with the highest response time.

In the event that the CUMP method produced a very low or very high threshold for an item (e.g., if the item was very easy or very difficult), two researchers independently visually examined the item’s response time frequency plot to determine a new threshold, using the VI method (Wise, 2006). The two researchers independently visually inspected the response time frequency plot, looking for a break point in which the distribution of rapid responses ends and the distribution of solution behavior responses begins (i.e., a dip between two response time modes). This visually-identified break point was used as the threshold.

Once determined, item response time thresholds were used to classify each item response, across all examinees and both time points, as either solution behavior or rapid responding. The proportion of total responses associated with solution behavior will determine each examinee’s response time effort (RTE; Wise & Kong, 2005) score (for each time point). The calculation of RTE is presented more formally in the equation below (adapted from Wise & Kong, 2005):
\[ SB_{ija} = \begin{cases} 
1 & \text{if } RT_{ija} > T_{ia} \\
0 & \text{if } RT_{ija} \leq T_{ia} 
\end{cases} \]  \hspace{1cm} (5)

\[ RTE_{ija} = \frac{\sum SB_{ija}}{k} \]  \hspace{1cm} (6)

Where \( SB \) indicates whether or not the response is classified as solution behavior, \( i \) indicates the item, \( j \) indicates the examinee, \( a \) indicates the time point, \( RT \) indicates the response time, \( T \) indicates the threshold, and \( k \) indicates the total number of items.

**Auxiliary variables.** Recall that auxiliary variables are included in MI for the primary purpose of increasing the possibility that the data will be MAR (rather than MNAR), and that in the context of this study, auxiliary variables should be related to either motivation-imposed missingness, or examinee quantitative and scientific reasoning ability, or both. Further, recall that if some solution-behavior responses are retained for a given examinee, and the items associated with these responses measure the same ability as any of the items with rapid-responses, these solution-behavior responses make ideal auxiliary variables. That is, the NW-9 item responses provided under solution behavior are ideal for controlling for a potential relationship between effort (missingness) and scientific and quantitative reasoning ability, and thus for increasing the likelihood of the MAR assumption over the MNAR assumption. In addition to variables included in hopes of inducing the MAR assumption, any variables that will be used for substantive analyses once MI has been performed should also be included during the MI process (Enders, 2010). Based on these criteria, and based on available data from students’ official university records, the following variables were included in the MI analyses in addition to the NW-9 solution behavior responses: SAT scores (math and verbal), GPA for Cluster 3
courses completed at post-test, and the number of Cluster 3 courses completed at post-test.

Although there are mixed research results pertaining to the relationship between effort and SAT scores (SAT scores have historically been used as a proxy variable for ability), I included these scores because I expect them to relate positively to examinee quantitative and scientific reasoning ability (measured via NW-9 scores). A positive relationship between SAT math and NW-9 scores has been empirically supported by Zilberberg, Finney, Marsh, and Anderson (2014); positive relationships between both SAT math and verbal scores and NW-9 scores have been alluded to by Hathcoat et al. (2015).

Applying a similar rationale as that for SAT scores, students’ GPA for Cluster 3 courses completed by the time of post-test was also included in the MI model. I hypothesize Cluster 3 course GPA at post-test will be positively related to NW-9 scores (as supported by Hathcoat et al., 2015; Sundre & Thelk, 2010), and may positively relate to effort.

The final auxiliary variable denotes the number of Cluster 3 courses an examinee has completed at the time of post-test. Examinees were classified into four groups, based on the number of relevant courses they have completed at the time of post-test: 0 courses, 1 course, 2 courses, and 3 or more courses (following the conventions of Hathcoat et al., 2015). Importantly, completing three of these relevant courses fulfills one of the General Education requirements at JMU, and so the ‘3 or more courses’ category is used both to represent those who have completed this requirement, and to avoid issues with sparseness of data associated with greater numbers of relevant courses completed (due to the limited
number of students who complete more than three relevant courses). This auxiliary variable was included for two reasons. First, I included it because of its relationship with NW-9 scores (as supported by Hathcoat et al., 2015; Sundre & Thelk, 2010). Second, I included it because it was used for substantive analyses following the MI procedure. These analyses are further described in the next section.

Data Analysis

Data analysis for this study occurred in three primary stages. First, three post-hoc methods for addressing examinee motivation were applied to the NW-9 data. This first step can be thought of as a type of data management. Second, the plausibility of the MCAR assumption was assessed. Finally, NW-9 scores were examined across the two time points. This third and final step includes the substantive analyses of interest.

Post-hoc methods for addressing examinee motivation. The three methods that were applied to the NW-9 data include (1) letting the data remain as they were observed (herein referred to as the unaltered method), (2) applying motivation filtering, and (3) imposing missingness at the item level and applying MI to create multiple, complete datasets.

The ‘unaltered’ method involves leaving the item responses untouched after they have been collected. This implies that the data to be analyzed include both solution behavior and rapid responding. Doing nothing is a fairly common method for analyzing data from low-stakes testing contexts. Practitioners may use this method because they either do not consider the impact of low examinee motivation on the quality of the data and on the inferences they draw from the data, they are not aware of the method of motivation filtering, they do not collect information about examinee motivation, or
perhaps simply because they do not wish to alter their data (e.g., by filtering). For this study, we applied this method by leaving the complete pre- and post-test NW-9 scores as they were observed, without regard to examinee motivation.

Motivation filtering, one of the most researched and applied methods for handling data collected in low-stakes testing environments, involves listwise deleting data associated with examinees categorized as having put forth low effort on their responses. For this study, examinees were categorized as ‘low effort’ for a given time point if their RTE scores for that time point were below .90, meaning that less than 90% of their responses were classified as solution behavior. The cutoff of .90 for RTE scores was selected based on its prevalent use in published research (e.g., Liu et al., 2015; Rios et al., 2014; Swerdzewski et al., 2011; Wise & Kong, 2005; Wise, 2015). Thus, examinee data were deleted listwise from the dataset if an examinee had a RTE score less than .90 at either time point (due to the necessity for a pre-post match).

The third and final method, MI, was used to impute multiple, plausible values for each rapid response in the dataset. Although both FIML and MI produce asymptotically equivalent results (Collins et al., 2001), MI was selected for this study due primarily to the ease with which it incorporates auxiliary variables and handles item-level missing data. Prior to applying MI, all scored item responses identified as rapid responses were converted to missing. That is, motivation-imposed missingness was applied to responses with associated response times lower than the item’s pre-determined response time threshold.

Once missingness was imposed, the imputation phase of MI was performed to create 100 complete datasets (i.e., $m = 100$). An $m$ of 100 was selected based loosely on
Graham et al.’s (2007) recommendation to impute many more than the standard three to five complete datasets, Enders’ (2010) recommendation to impute a minimum of 20 complete datasets, and Johnson and Young’s (2011) recommendation to impute more than 25 complete datasets. For the imputation phase of MI, a non-informative prior was used to estimate posterior distributions for the regression coefficients.\textsuperscript{11}

In an effort to avoid issues with non-convergence of the posterior distributions, high burn-in and thinning values were used for the MI process.\textsuperscript{12} Specifically, 10,000 burn-in iterations and 10,000 thinned iterations were used. That is, the first 10,000 iterations were discarded, and 10,000 iterations between each kept iteration were discarded. Thus, a total of 1,00,100 iterations were completed in order to arrive at $m=100$ complete datasets.

Recall that the imputation phase of MI includes using stochastic regression procedures to impute plausible response values for missing data points. In this study, between 5 and 135 variables served as predictors in the stochastic regression equations. For example, for examinees who only rapidly responded to one item across both time points, all 131 (65 at one time point + 66 at the other = 131) solution behavior item responses and all 4 auxiliary variables were used to predict plausible values for that item for those examinees. On the other extreme, if some examinees rapidly responded to all

\textsuperscript{11} This is the standard prior specification for single-level multiple imputation within the Blimp software program (see Keller & Enders, 2017).

\textsuperscript{12} Ideally, convergence would be assessed visually (e.g., via trace plots) and using diagnostic indices (e.g., potential scale reduction (PSR) factor (Gelman & Rubin, 1992)); however, neither of these options were feasible for this study. First, the software used for imputation, Blimp (Enders et al., 2016; Keller & Enders, 2017), does not currently provide visual convergence inspection options. Second, due to the large number of parameters being estimated, an overwhelming number of PSR values would need to be generated and evaluated. Ultimately it was the advice of one of the software’s authors that we address convergence via large burn-in and thinning values, rather than attempt to garner and evaluate PSR factors, for this study (C. Enders, personal communication, February 24, 2017).
but one item across both time points, the one solution behavior response and the 4 auxiliary variables were used to predict plausible values for each of their 131 missing responses. Note that a maximum of 135 auxiliary variables would ideally be accompanied by a dataset with at least 405 complete cases, to be in accordance with Hardt et al.’s (2012) preliminary recommendation of one auxiliary variable per a minimum of three complete cases (135 x 3 = 405).

Item responses were treated as ordinal variables in the regression model, and thus a probit regression model was used to predict the imputed values in the Blimp software (Enders, Keller, & Levy, 2016; Keller & Enders, 2017). Further, because item responses were treated as ordinal, the imputed values took on whole numbers (sidestepping the issue of whether or not to round the values; see Wu, Jia, & Enders, 2015).

Assessing the plausibility of the MCAR assumption. Because motivation filtering relies on the assumption that the data are MCAR, prior to filtering out low-motivated-examinee data comparisons of solution-behavior item scores between low- and adequately-motivated examinees, as well as standard tests for the MCAR mechanism, were performed. This was done to assess the plausibility of the MCAR assumption (research question one).

Although motivation filtering will remove all of the item responses for examinees with RTE scores less than .90, I believe that some of these item responses, the solution-behavior responses, represent viable data. If I examine the portion of solution-behavior responses that are correct from examinees with RTE scores less than .90, and compare them to the proportion of solution-behavior responses that are correct for examinees with RTE scores greater than or equal to .90, I can examine an approximation for the
relationship between effort and ability. If the data were MCAR, there would be no
relationship between effort and ability, and thus I would expect no relationship between
effort and scores associated with solution behavior (a proxy for ability). If the data were
MNAR (which I would hope to move to MAR using the solution-behavior responses as
auxiliary variables in conjunction with MI), there would be a relationship between effort
and ability, and thus I would expect a relationship between effort and scores associated
with solution behavior.

Recall that a test of MCAR involves testing for relationships between missingness
(i.e., whether or not an examinee’s data will be filtered) and other measured variables
(e.g., SAT scores). If relationships are found between missingness and observed
variables, I can rule out the presence of the MCAR mechanism. Although I do not believe
the MCAR mechanism is plausible for these data, testing MCAR serves two purposes: (1)
to identify relationships between missingness and other measured variables (and thus
further justifying the use of the auxiliary variables), and (2) given the results support
relationships between missingness and other variables, to demonstrate the
inappropriateness of listwise deletion.

Substantive analyses. Across each of the three methods (i.e., unaltered,
motivation filtering, and MI), pre- and post-test NW-9 scores were examined. Pre-post
group mean differences were examined to determine if the conclusions and inferences we
make about student growth in quantitative and scientific reasoning ability depend on how
we treat rapid responses. Due to the applied nature of this study, however, it is not
possible to definitively determine which method produced the most accurate growth
scores. To supplement any findings indicating an effect of method, we also examined
relationships between change over time and other variables. Specifically, we investigated the relationship between mean differences in NW-9 scores and the number of relevant courses completed at post-test—a relationship commonly examined by assessment practitioners at our university (Hathcoat et al., 2015; Pieper, Fulcher, Sundre, & Erwin, 2008; Sundre & Thelk, 2010). Based on previous research, we expected growth in NW-9 scores to be positively related to the number of relevant courses completed. Thus, we expected methods that accurately estimate growth scores to also show a positive relationship between mean differences in NW-9 scores and the number of relevant courses completed.

The results from each method were first compared via descriptive statistics for NW-9 pre-test, post-test, and difference scores. Means, standard deviations, and standard errors of the means were examined. Following an analysis of descriptive statistics, a mixed ANOVA (i.e., a split-plot ANOVA) was used to investigate the effects of time and the number of relevant courses completed on NW-9 scores. To do this, NW-9 scores were predicted by time (two levels: pre-test, post-test), number of courses completed (categorized into four levels: 0, 1, 2, 3+ courses), and the interaction between time and number of courses completed.

In order to obtain the necessary parameter estimates for these substantive analyses, the mixed ANOVAs were modeled within the regression framework and estimated using maximum likelihood estimation. First, the variables of time, number of courses, and their interaction were effect coded into one, three, and three variables, respectively (as recommended by van Ginkel & Kroonenberg, 2014). Then, a multiple regression model with the seven effect-coded predictors predicting NW-9 scores was
performed using the PROC MIXED command in SAS. To assess the significance of the main effects and interaction from these models, multiparameter significance tests were performed. For the unaltered and motivation filtered methods, multivariate Wald tests were performed. Essentially, these tests assess the significance of multiple parameters as a group. For example, to assess the significance of the effect of number of courses, the three effect-coded parameters were simultaneously tested against a null hypothesis that they equal zero. For the MI method, the $D_1$ multiparameter test (Li, Raghunathan, & Rubin, 1991) was performed. The $D_1$ test is analogous to a Wald test, and has been recommended for use with pooled ANOVA parameter estimates (e.g., Grund, Lüdtke, & Robitzsch, 2016). Further detail regarding the process of obtaining parameter estimates and their associated multiparameter significance tests for each of the post-hoc methods, using SAS, is provided in Appendix A.

A significant main effect of time would indicate that NW-9 scores at pre-test were significantly different from those at post-test. A significant main effect of number of courses completed would indicate that NW-9 scores significantly differed depending on the number of courses completed. A significant interaction would indicate that the change in NW-9 scores from pre- to post-test depended on the number of courses completed. Previous research suggests that each of these effects should be present in the data. That is, we expected NW-9 scores to increase from pre- to post-test, to be greater for those with more courses completed, and to increase more for those who complete more courses (as supported by Hathcoat et al., 2015; Sundre & Thelk, 2010).

Statistical significance was assessed for each of the three effects in the mixed ANOVA model (i.e., time, number of courses, interaction between time and number of
courses). Specifically, $F$ statistics and $p$ values, were compared across rapid response treatment methods (addressing research questions three and four).

For the unaltered and motivation filtering methods, these analyses were performed just as they would on any other complete dataset. For the MI method, these analyses were performed on each of the 100 imputed datasets separately, and then the 100 sets of results were pooled to create one final set of results (the analysis and pooling phase of MI). Although the pooled results from the 100 imputed datasets were the primary focus of this study, subsamples of these 100 imputed datasets were also pooled for comparison. Specifically, results from the first 3, 5, 10, 25, and 50 imputed datasets were pooled. This was done to examine the ‘final’ pooled estimates as a function of the number of imputed datasets (i.e., research question two).

**Software**

Data management was performed using SPSS version 23. CUMP thresholds were set using R version 3.2.1 (R Core Team, 2015). See Appendix B for R syntax. The imputation portion of the MI procedure was performed using beta version 6.71 of the Blimp software program (Enders et al., 2016; Keller & Enders, 2017). See Appendix C for Blimp syntax. Substantive data analyses described above (including pooling parameter estimates and standard errors across imputations) were performed using SAS version 9.4. See Appendix D for SAS syntax.
CHAPTER FOUR

Results

Introduction

In the following chapter I present the results of this study. First, I describe the
data associated with each of the measures (e.g., NW-9 scores, response times). Second, I
present the results of applying the three post-hoc methods to the observed NW-9 scores.
Third, I present the results pertaining to the assessment of the plausibility of the MCAR
assumption (addressing research question one). Finally, I present the results of the
substantive analyses (addressing research questions two, three, and four).

Description of Data

In this section I describe the data for each of the measures used in this study.
Specifically, I describe the observed NW-9 responses, NW-9 response times, NW-9
response time thresholds, solution behavior and rapid response rates for the NW-9
responses, resulting RTE scores, SAT math and verbal scores, Cluster Three GPA at
post-test, and number of Cluster Three courses completed at post-test.

**Observed NW-9 responses.** Examinees’ observed (scored) responses to the 66
NW-9 items were summed to create total scores at pre- and post-test. On average,
examinees scored 44.98 out of 66 (68.15% correct) at pre-test, with about 68% of the 388
examinees scoring between 38.34 (58.08% correct) and 51.63 (78.22% correct) at pre-
test. Observed pre-test scores were approximately normally distributed (skewness = -
0.37, kurtosis = -0.02). At post-test, on average, examinees scored 48.79 out of 66
(73.92% correct), with about 68% of the 388 examinees scoring between 41.61 (63.04%
correct) and 55.97 (84.81% correct) at post-test. Observed post-test scores were approximately normally distributed (skewness = -0.63, kurtosis = 0.19).

**Examinee effort.** Examinee effort was operationally defined according to examinees’ observed response times for each of the 66 NW-9 items at pre- and post-test. Item response times tended to be positively skewed and leptokurtic, for both pre- and post-test. For example, the distribution of response times for item 5 at post-test is shown in Figure 1. Item response time distributions varied by item. Because item response times were not normally distributed, each item’s median response time was used as a summary statistic. Median item response times ranged from 5 to 85 seconds, with post-test medians generally lower than pre-test medians. See Figure 4.

Item response time thresholds were set using both the CUMP and VI methods. The CUMP method successfully identified thresholds for 40 of the 66 items at pre-test, and 41 of the 66 items at post-test. When the CUMP method was unable to set thresholds, it was typically because the relative cumulative proportion of examinees’ correct responses across time was always greater than the chance correct value (that is, the majority of examinees responded correctly, even at low response times), although for a few items it was because the relative cumulative proportion of examinees’ correct responses was always less than the chance correct value (that is, the majority of examinees responded incorrectly).

Two raters, the chair of this dissertation and I, provided thresholds for the items without CUMP thresholds using the VI method. Table 5 denotes which threshold-setting method was used for each item for each time point. The raters’ thresholds were within two seconds of each other for 78% of the ratings. If ratings were more than two seconds
apart for any given item, item response time and item response accuracy data were reviewed and a final threshold determination was made. The final item response time thresholds, including both CUMP and VI thresholds, ranged from 0.5 to 43 seconds. Each item’s response time threshold is displayed in Figure 5 and Table 5. Although item response time thresholds did not have an overall tendency to increase or decrease from pre- to post-test (average difference from pre- to post-test was -0.1 seconds), the pre- and post-test thresholds were positively correlated \( r = 0.64 \).

The item response time thresholds were then used to determine whether examinees exhibited solution behavior or rapid responding for each item (for each examinee). At the item level (i.e., examining the proportion of examinees who exhibited solution behavior for a given item), generally, over 90% of examinees exhibited solution behavior. At pre-test, four items had less than 90% of examinees exhibit solution behavior (i.e., items 39, 53, 62, and 63). At post-test, five item had less than 90% of examinees exhibit solution behavior (i.e., items 2, 10, 61, 62, and 63). The items with the lowest percentage of solution-behavior responses were item 63 at pre-test (51% solution behavior), and item 2 at post-test (35% solution behavior). At the examinee level (i.e., examining the proportion of items for which an examinee exhibited solution behavior), generally, RTE scores were high. At pre-test, RTE scores ranged from .41 to 1.00, with a median RTE score of .97. At post-test, RTE scores ranged from .30 to 1.00, with a median RTE score of .97. See Figure 6.

\[\text{Collectively, these seven items were examined with regard to item characteristics (i.e., difficulty, tendency to be part of a testlet, and stem word count). The items tended to be fairly difficult (see Table 5), with the exception of item 62. Six of these items were part of testlets (e.g., items 60-63 formed a testlet; item 2 was not part of a testlet. The seven items varied in stem length (e.g., item 10’s stem had 11 words, item 61’s stem had 95 words), and in whether or not a graphic was part of the item.}\]
**Auxiliary variables.** In addition to the solution-behavior item responses, four auxiliary variables were used in this study, namely, SAT math scores, SAT verbal scores, Cluster Three GPA at post-test, and the number of Cluster Three courses completed by post-test.

For the 373 examinees for whom SAT math scores were available, SAT math scores were normally distributed with a mean of 584.43 on a scale of 200-800 possible points ($SD = 65.86$). For the 371 examinees for whom SAT verbal scores were available, SAT verbal scores were normally distributed with a mean of 573.46 out of a possible 200-800 points ($SD = 68.42$).

On average, examinees had a Cluster Three GPA of 2.94. Cluster Three GPA scores were approximately normally distributed, with about 68% of examinees having a Cluster Three GPA between 2.17 and 3.70. These GPA scores were only based on examinees who had completed Cluster Three coursework by post-test, which was the majority (91.5%), but not the entirety, of the sample. The distribution of the number of Cluster Three courses completed by post-test is displayed in Table 6.

**Applying Post-Hoc Methods**

Once the item response time thresholds and RTE scores had been determined, the three post-hoc methods for addressing examinee motivation were applied. The first method, leaving the data as they were observed, simply retained the entire dataset of observed responses to the NW-9 items at pre- and post-test ($N = 388$).

Motivation filtering was applied by listwise deleting any examinee who had an RTE score less than .90 at *either* time point. For the pre-test data, 32 examinees (8.2% of the sample) had RTE scores less than .90. For the post-test data, 34 examinees (8.8% of
the sample) had RTE scores less than .90. Of these 34 examinees, only 13 also had RTE scores below .90 at pre-test. This means that 19 of the 32 examinees with RTE scores below .90 at pre-test had RTE scores equal to or above .90 at post-test, and that 21 of the 34 examinees with RTE scores below .90 at post-test had RTE scores equal to or above .90 at pre-test. That is, 19 examinees displayed low effort at pre-test, yet displayed adequate effort later on at post-test, and 21 examinees displayed adequate effort at pre-test, yet displayed low effort later on at post-test (and 13 examinees displayed low effort at both time points). When motivation filtering was applied across both time points, a total of 53 examinees were removed from the dataset (n = 335).

The first step in applying the third post-hoc method, MI, was to impose missingness at the item level for any item response not associated with solution behavior (i.e., rapid responses). Thus, first, all rapid responses (across both time points), were set to missing. This resulted in a loss of 4.2% of the data. That is, out of a possible 51,216 item-examinee pairs (51,216 = 388 examinees * 66 items * 2 time points), 49,080 (95.8%) were retained and 2,136 (4.2%) were set to missing. This included 309 unique patterns of missingness within examinees. The most common pattern of missingness contained observed data for all variables (including SAT scores and Cluster Three information) except for item 2 at post-test; 22 (5.7%) of the 388 examinees had this pattern of missingness. The second most common pattern of missingness was to have no missing data for any of the variables; 17 (4.4%) examinees had this pattern of missingness. That is, there were 17 examinees with complete data for all variables in this dataset (at both time points). The majority of the sample (59%) had 4 or fewer missing
values across all 136 variables. See Table 7 for a summary of the amount of missingness within examinees in this sample.

MI was then applied to the missing-imposed data, using the solution-behavior responses, SAT scores, Cluster Three GPA, and number of Cluster Three courses completed as auxiliary variables. The MI model, with 10,000 burn-in iterations and a thinning parameter of 10,000 iterations, produced the 100 imputed datasets after 234 hours and 10 minutes (approximately 10 days). The 100 imputed datasets were then each ready to be analyzed ($N = 388$, for each dataset).

**Assessing the Plausibility of the MCAR Assumption**

The plausibility of the MCAR assumption was assessed by comparing low- and adequately-motivated examinees’ scores across the different auxiliary variables. Specifically, we used independent-groups $t$-tests to compare these two groups of examinees on their solution-behavior-item-only percent correct scores, SAT scores, Cluster Three GPA, and the number of Cluster Three courses they had completed. Low-motivated examinees were identified by RTE scores less than .90 (i.e., examinees who would be filtered out of the dataset). Adequately-motivated examinees were identified by RTE scores equal to or greater than .90. The results of these comparisons are provided in Table 8.

For the pre-test data, the 32 examinees with RTE scores less than .90 had an average percent correct score (again, based only on their solution-behavior responses) of .61 ($SD = .11$), and the 356 examinees with RTE scores equal to or greater than .90 had an average percent correct score (based only on their solution-behavior responses) of .70 ($SD = .09$). These means were statistically and practically significantly different.
from one another ($M_{\text{difference}} = .10, p < .001, \text{Cohen’s } d = 0.96$). Thus, at pre-test, low-motivated examinees’ average percent correct score (based only on solution-behavior responses) was statistically and practically significantly lower than adequately-motivated examinees’ average percent correct score (based only on solution-behavior responses).

For the post-test data, the 34 examinees with RTE scores less than .90 had an average percent correct score (based only on their solution-behavior responses) of .67 ($SD = .09$), and the 354 examinees with RTE scores equal to or greater than .90 had an average percent correct score (based only on their solution-behavior responses) of .76 ($SD = .10$). These means were statistically and practically significantly different from one another ($M_{\text{difference}} = .09, p < .001, \text{Cohen’s } d = 0.98$). Thus, at post-test, low-motivated examinees’ average percent correct score (based only on solution-behavior responses) was statistically and practically significantly lower than adequately-motivated examinees’ average percent correct score (based only on solution-behavior responses).

Because the other auxiliary variables’ scores (i.e., SAT scores and Cluster Three information) are time invariant in this study, group comparisons for these variables were conducted by comparing the 53 examinees with RTE scores lower than .90 at either time point to the 335 examinees whose RTE scores were equal to or above .90 for both time points. Recall that not all examinees had SAT score data, though, so for these comparisons the sample sizes for the two groups do not add up to the full 388 examinees. Average SAT math scores did statistically and practically significantly differ between groups, with the adequately-motivated examinees having a higher average SAT math score than the low-motivated examinees ($p = .024, \text{Cohen’s } d = 0.34$). Average SAT verbal scores, on the other hand, did not statistically significantly differ between groups.
Further, no statistically significant differences were found between groups for Cluster Three GPA or number of Cluster Three courses completed.

**Substantive Analyses**

The primary, substantive analysis of interest for this study was a mixed ANOVA assessing the main effects of time and number of Cluster Three courses completed at post-test, and their interaction, on examinees’ NW-9 scores. Prior to conducting this mixed ANOVA, though, descriptive statistics for the NW-9 scores for each of the three post-hoc methods were examined (Tables 9-13; Figure 7). For the MI method, this meant that descriptive statistics were calculated for each of the \( m = 100 \) imputed datasets, and then the results were pooled.

In Tables 9-13 and Figure 7 we see a slight increase in NW-9 scores over time for each method, indicating a potential positive main effect of time on NW-9 scores across methods. In Tables 9-11 we see this by comparing the marginal values for pre- and post-test (column marginals) for each method; in Figure 7 we see this by comparing the pre-(bottom) and post-test (top) lines for each method; in Tables 12-13 difference scores are displayed explicitly. It does not appear that there is much potential for a main effect of number of Cluster Three courses completed for any of the methods, however. We see this in the little variation in marginal values for the number of Cluster Three courses (row marginals) in Tables 9-11, and in the little change of the shapes of the lines across the number of Cluster Three courses (across the x-axis) for each method in Figure 7. To examine the descriptives for a potential interaction effect, we compare the change in pre- and post-test scores for each of the groups of Cluster Three course completion, for each method, in Table 13. Alternatively, we can compare the distance between corresponding
pre- and post-test lines across each of the numbers of Cluster Three course completion in Figure 7. For example, based on Figure 7, it appears that the distance between the red solid lines is greater for those who have completed three or more Cluster Three courses, compared to those who have completed two courses. We must note, however, that the range of the y-axis of Figure 7 does not reflect the full range of NW-9 scores, and thus may distort our interpretation of the magnitude of the effects. To account for this, Figure 8 displays the same data as Figure 7, albeit with the full range of possible NW-9 scores. Figure 8 seems to indicate that we may not find a significant main effect of number of Cluster Three courses, nor a significant interaction.

Aside from foreshadowing the results of the ANOVA, we should also examine the descriptive statistics by post-hoc method. When we compare NW-9 scores by post-hoc method, we see that the unaltered method consistently produced the lowest mean scores for pre-test, post-test, and change from pre- to post-test, the motivation filtering method consistently produced the highest mean NW-9 scores, and that the pooled MI scores were consistently higher than those from the unaltered data and lower than those from the motivation-filtered data. The exact opposite pattern was true for the standard deviation of NW-9 scores by method: the largest standard deviations were associated with the unaltered data, then the MI data, and the smallest standard deviations were associated with the motivation-filtered data.

**Mixed ANOVA results.** Mixed ANOVAs were performed for each of the three post-hoc methods to assess the main effect of time, the main effect of the number of Cluster Three courses completed by post-test (categorized into four groups), and the interaction of time and number of Cluster Three courses completed by post-test on
examinees’ NW-9 scores. Prior to performing the mixed ANOVAs, assumptions related to the normality of NW-9 scores and structure of (co)variance matrices were assessed and were found to be satisfied.\textsuperscript{14} For the MI data, mixed ANOVAs were performed on each of the $m = 100$ imputed datasets, and the results were pooled. Table 14 displays the resulting parameter estimates and standard errors for each of the parameters from the mixed ANOVAs. Although the parameter estimates varied by method, there was no overall pattern to the variation (i.e., no method consistently produced the highest or lowest parameter estimates). There was a pattern for the standard errors, though; the unaltered data condition consistently produced the largest standard errors for all parameters. The motivation filtered and MI data produced similar standard errors for each parameter; the standard errors for these methods were within 0.01 units of one another for each parameter. Neither of these methods (motivation filtered and MI) consistently produced the smallest standard errors.

The parameter estimates from Table 14 were used to perform multivariate Wald and $D_1$ tests to assess the significance of the two main effects (i.e., time and number of Cluster Three courses completed at post-test) and their interaction. Table 15 contains the results of these tests. Generally, the conclusions regarding the statistical significance of

\textsuperscript{14} Normality of the NW-9 scores was assessed through examination of skewness and kurtosis values for each combination of levels of the two independent variables (i.e., time and number of courses). The structure of the error covariance matrix was assessed by applying a compound symmetric structure to the error covariance matrix and comparing the fit of this model to that when an unstructured covariance matrix was applied. The compound symmetric structure assumes equal variances for pre- and post-test scores both within a group and across the different groups of course completion; it also assumes equal covariances across groups of course completion status. The unstructured matrix allowed the variance of scores from pre-test to differ from post-test both within and across groups; it also allowed covariances to differ across groups. For each of the 102 datasets analyzed in this study ($102 = 100$ imputed datasets + 1 unaltered dataset + 1 motivation-filtered dataset), the information criteria (i.e., AIC and BIC) values were consistently lower (indicating better fit) when compound symmetry was assumed, compared to when the unstructured matrix was applied. This indicated that the added complexity of the unstructured covariance matrix did not provide substantially better fit than the compound symmetric matrix. Thus, a compound symmetric matrix was applied.
the effects hold for each of the post-hoc methods; there was no significant interaction
effect,\textsuperscript{15} nor was there a significant main effect of number of courses. There was a
significant main effect of time. Within each effect, the test statistic values (i.e., $F$ values)
varied by post-hoc method, yet not consistently. For the (non-significant) interaction
effect, the motivation filtering method produced the most extreme test statistic, and the
MI method produced the least extreme test statistic. For the (non-significant) main effect
of number of courses, the MI method produced the most extreme test statistic, and the
motivation filtering and unaltered methods produced similar (less extreme) test statistics.
Finally, for the significant main effect of time, the motivation filtering method produced
the most extreme test statistic, and the unaltered method produced the least extreme test
statistic.

**Comparing MI results by number of imputed datasets.** Although the pooled
results from the analyses of the 100 multiply-imputed datasets were the primary focus of
this study, subsamples of these 100 imputed datasets were also pooled for comparison.
Table 16 displays descriptive statistics for NW-9 scores pooled from the first 3, 5, 10, 25,
and 50 imputed datasets. Table 17 displays the mixed ANOVA parameter estimates and
standard errors, pooled from the first 3, 5, 10, 25, and 50 imputed datasets. Finally, Table
18 displays the significance test results for the main effects and interaction effect of the
mixed ANOVA, pooled from the first 3, 5, 10, 25, and 50 imputed datasets. Generally,
the parameter estimates and standard errors showed little change across the number of
imputed datasets. Tables 19 and 20 contain between- and within-imputation variance
information for each number of imputed datasets. Between- and within-imputation

\textsuperscript{15} In comparison to the standard nominal alpha of 0.05.
variance information was included in effort to better understand the patterns (or lack thereof) in pooled standard errors across numbers of imputations.
CHAPTER FIVE

Discussion

Introduction

Generally, the role of this chapter is to discuss the results of this dissertation in the context of (a) the purposes of this dissertation, and (b) application to real-world practice. To do this, I begin by reviewing the two purposes of this dissertation. Then, I summarize the empirical results for each of the four research questions, in turn. Following this, I discuss recommendations for researchers and practitioners, primary limitations of this study, and ideas for future research. This chapter ends with an overall summary of this dissertation.

Purposes of this Dissertation

This dissertation served two purposes. The first purpose, which was didactic in nature, was to reframe current post-hoc methods for addressing examinee motivation as missing data methods. This first purpose was addressed by making explicit connections between post-hoc approaches for addressing examinee motivation and missing data techniques. For example, I discussed how motivation filtering can be conceptualized as a form of listwise deletion. The intention of re-framing post-hoc methods for addressing examinee motivation as missing data methods was to help practitioners think about the appropriate (and inappropriate) uses of these methods for handling data impacted by low examinee motivation. For example, although it has been suggested that practitioners adopt the use of motivation filtering (e.g., Wise & DeMars, 2010), it is rarely made explicit that motivation filtering assumes the data are filtered completely at random (i.e., are MCAR). By thinking about motivation filtering as a missing data technique (i.e.,
listwise deletion) practitioners may be more likely to acknowledge that there are specific assumptions, and thus specific appropriate uses, of this method.

In addition to demonstrating connections between post-hoc approaches for addressing examinee motivation and traditional missing data techniques (e.g., listwise deletion), I also discussed connections to the ‘modern’ missing data techniques (i.e., FIML, MI). For example, the effort-moderated IRT model (Wise & DeMars, 2006), which treats rapid responses as if they were missing, can be conceptualized as a form of FIML. Further, although it has seen little use, MI can also be used to treat rapid responding. Like the effort-moderated IRT model, using MI to treat rapid responding requires imposing missingness in place of rapid responses. Some advantages of using a modern missing data method as a post-hoc treatment of rapid responding are that it retains the entire sample of data, and that it produces unbiased parameter estimates and standard errors when the data are MAR (or MCAR). This is in contrast to the traditional methods, like motivation filtering, which do not retain the entire sample of data, and only provide unbiased estimates when the data are MCAR. The difference between MCAR and MAR is crucial in this context because motivation-imposed missing data are not likely to be MCAR. In fact, some research suggests they may be MNAR (e.g., Rios et al., 2017; Wise et al., 2009). Luckily, modern missing data methods have another advantage over traditional methods: they can incorporate auxiliary variables, which serve to transition the missingness mechanism from MNAR to MAR.

In the context of motivation-imposed missingness, the best auxiliary variables are those that relate to the values of the variable(s) with missing data. This is in contrast to typical recommendations, which suggest the use of auxiliary variables that relate to either
the values of the missing variable(s), or the missingness, or both. In this context, however, we do not want to use information about examinees’ motivation (i.e., information about missingness) because we wish to impute their scores had they put forth good effort; we do not wish to impute scores associated with low-effort. Thus, in this context, the best auxiliary variables are those that relate to the values of the items with missing data. Fortunately, when we impose missingness at the item level we have ideal auxiliary variables in the non-missing item responses (i.e., in the solution-behavior responses). Because we assume the non-missing responses are indicative of the examinee’s ability on the construct measured by the assessment, we can use these responses to essentially control for a relationship between missingness and the values of the missing item responses. That is, by using the solution-behavior item responses as auxiliary variables we believe we are able to transition the data from MNAR to MAR, rendering the modern missing data methods appropriate for application.

In sum, the first purpose of this dissertation was fulfilled by discussion of post-hoc approaches to handling low examinee motivation as traditional and modern missing data methods. This discussion included direct links from specific approaches to specific methods, as well as note of when traditional vs. modern methods are appropriate for use. Although the focus of this dissertation was on low-stakes testing situations, the re-framing of rapid-responses as missing responses is also applicable to high-stakes testing situations. In these instances, however, the rapid-responses are less likely to indicate low motivation, and instead are more likely to indicate a speeded assessment.

The second purpose of this dissertation, which was empirical in nature, was to compare three post-hoc methods for addressing low examinee motivation using real data.
Specifically, I compared the results of substantive analyses based on leaving the data unaltered, motivation filtering (listwise deleting) the data, and multiply imputing the data. For this second purpose of the dissertation, four research questions were addressed. The results are reviewed next, by research question.

**Summary of Results by Research Question**

The empirical portion of this dissertation aimed to address four research questions. In the following section I summarize the results for each research question and consider the extent to which the results align with expectations.

**Research question one.** The first research question addressed the plausibility that the data are MCAR for the motivation filtering method. When we apply motivation filtering, we are assuming the data are MCAR. If this assumption is violated, it means that our resulting parameter estimates are likely biased, and thus the validity of the inferences we make about examinees’ proficiency or learning (based on these parameter estimates) is questionable. The plausibility of the MCAR assumption was assessed by comparing low- and adequately-motivated examinees’ scores for several variables. Because examinees were filtered, or listwise deleted, based on their designation as demonstrating either low (RTE < .90) or adequate (RTE ≥ .90) test-taking effort, comparisons between these two groups serve as comparisons between those with missingness and those without. Recall that for the MCAR assumption to hold, missingness cannot relate to the values of the missing variables themselves (a relationship that is not *directly* testable), nor can it relate to any other measured variable in the dataset (relationships that are directly testable). Investigating the presence of these relationships was the goal of the first research question. Specifically, research question one asked,
How do solution-behavior item scores and other auxiliary variable scores from examinees demonstrating adequate test-taking effort (i.e., those with RTE scores ≥ .90) compare to those from examinees demonstrating low test-taking effort (i.e., those with RTE scores < .90)?

Regarding group comparisons for NW-9 scores based only on solution-behavior responses, in Table 8 we saw that adequately-motivated examinees had statistically and practically significantly higher NW-9 solution-behavior-item scores at both pre- and post-test. Although this is not a direct assessment of a relationship between the missing responses and missingness, these results suggest a relationship may be present. That is, although we could not assess the relationship between low- and adequately-motivated examinees’ total NW-9 scores (because examinees did not display solution behavior for all items), we could assess differences in a proxy for these scores (i.e., examinees’ scores based solely on their solution-behavior responses), and when we did, we found a relationship. On average, the examinees who were motivation filtered had a lower percent correct score for the NW-9 items to which they responded using solution behavior, compared to the examinees who were not motivation filtered. These results were true for both pre- and post-test scores. Importantly, these results indicate that when motivation filtering is applied to these data, lower-ability examinees are being removed from the dataset and thus from subsequent substantive analyses. These results indicate that the data are not likely MCAR.

Regarding group comparisons for the other auxiliary variables, in Table 8 we saw that adequately-motivated examinees had statistically and practically significantly higher SAT math scores, compared to low-motivated examinees. This relationship provides
further indication of a relationship between motivation and quantitative (and scientific) reasoning ability.

Motivation groups did not differ, however, on SAT verbal, Cluster Three course GPA, or number of Cluster Three courses completed by post-test. These results present a direct test of the MCAR assumption, and indicate that there is a relationship between missingness and a measured variable in the dataset (i.e., SAT math scores).

Ultimately, given the assumption of MCAR is violated (meaning that the data are either MAR or MNAR), we expect that the parameter estimates resulting from the motivation filtering method will be inaccurate, and thus that the inferences we would make based on these parameter estimates would not accurately reflect our population of interest. If the data are MNAR, though (as suggested by these results), we would also expect inaccurate parameter estimates to result from the MI method, unless we could control for the suggested relationship between ability and effort through the use of auxiliary variables (thus rendering the data MAR). Of critical importance to this study is that because we used the solution-behavior item scores and other variables related to quantitative and scientific reasoning ability (e.g., SAT math) as auxiliary variables for the MI method, we made an effort to achieve this. Thus, for these data, the results of research question one indicate that of the methods studied here, the only appropriate post-hoc method for handling data from low-motivated examinees was to use MI with solution-behavior item scores and other variables related to quantitative and scientific reasoning ability as auxiliary variables.16

---

16 Because MI and FIML are asymptotically equivalent methods, FIML with solution-behavior item scores and other variables related to quantitative and scientific reasoning ability as auxiliary variables would also be appropriate (although this method was not studied here).
**Research question two.** The second research question pertained only to the MI method. More specifically, it pertained to the results obtained using MI as the number of retained imputations ($m$) increased. The second research question asked, How do the pooled standard errors compare across different numbers of imputations for MI? The results of this research question are intended to add to the limited pool of research regarding the appropriate number of imputations to retain when using MI.

In Table 16 we saw that as the number of imputations increased, in general the pooled standard errors tended to decrease slightly. The largest relative decrease in standard errors occurred between $m=3$ and $m=5$, with smaller relative decreases occurring beyond $m=5$. This is a similar pattern to what Graham et al. (2007) found (i.e., that the pooled standard errors decreased slightly and then began to settle as the number of imputations increased from 3 to 100), although we did not see the standard errors settle as much in this study. In fact, some standard errors increased as number of imputations increased. For the standard errors in Table 17, however, a pattern in which standard errors decreased and leveled off as number of imputations increased did not hold. That is, we did not see an overall decrease in the standard errors as the number of imputations increased. This trend was present for the parameter associated with time, but not for others. The standard errors associated with the other parameters tended to be consistent, or even to increase slightly, as the number of imputations increased. These results are contrary to what we would expect. Because $m$ is included in the equation for calculating pooled standard errors, we would expect the pooled standard errors to decrease as a function of $m$. 


In an effort to diagnose the misalignment between these results and expectation, the between- and within-imputation variances of the parameter estimates were examined (recall that these two types of variances are combined to create the pooled standard errors). In Table 19 we see that the between-imputation variance tends to decrease as $m$ increases, and the within-imputation variance varies but not systematically. In Table 20 we do not see this pattern for the between-imputation variance (except for the time and two courses variables). These results indicate that it is not necessarily true that pooled standard errors will always decrease as a function of the number of imputed datasets.\footnote{One might argue that the trends in these data are over interpreted here, or that we are “splitting hairs” regarding the number of places after the decimal we are looking to examine differences. This may be the case. However, we note that Graham et al. (2007) interpreted differences in pooled standard errors that occurred in the fourth decimal place.}

Further, these results indicate that regardless of the size of $m$ the same substantive conclusions would be drawn. That is, for this study no matter how many imputations were analyzed the results consistently indicated a statistically significant main effect of time, and non-statistically significant effects of number of Cluster Three courses completed and time by number of courses interaction.

Although unfortunately the results associated with research question two seem to provide little guidance as to the number of imputations to retain when using MI, we recognize Enders’ (2010) recommendation to impute a minimum of 20 datasets. Further, we note that additional imputations beyond $m=20$ may be beneficial for decreasing pooled standard errors and thus for increasing power for statistical significance testing. It is up to the researcher whether this potential benefit is worth the resources to impute many more datasets.\footnote{Recall that obtaining 100 imputations took approximately 10 days for this study.}
**Research question three.** The third research question addressed whether our inferences about pre-post growth in quantitative and scientific reasoning would differ according to the method used for addressing low examinee motivation. Specifically, research question three asked, How do the statistical significance conclusions regarding the effect of time on NW-9 scores compare across the three post-hoc methods for handling examinee motivation (i.e., leaving the data intact, motivation filtering, and using multiple imputation to impute plausible replacement values for rapid responses)?

Because previous results (within this study) suggested a positive relationship between ability and effort, we expected the parameter estimates from motivation filtered data to be greater than those from the observed (unaltered) data. For example, estimates of the overall population’s average scientific and quantitative reasoning ability were expected to be greater than those from the unaltered data because they would tend to be based on the moderate-to-high ability examinees’ NW-9 scores. If the data were effectively MAR due to the use of the solution-behavior item auxiliary variables with the MI method, we would expect the parameter estimates from the MI method to be lower than those from the motivation filtered data. The unaltered data were expected to produce the lowest estimates of ability because this method retained and analyzed rapid responses, which tend to have only a chance level of being correct. In sum, we expected that the unaltered data would produce the lowest estimates, the MI data would produce the next highest estimates, and that the motivation filtered data would produce the highest estimates of ability. This anticipated pattern of estimates has been supported by previous research when ability is related to motivation (e.g., Koepfler et al., 2011), as is plausible
here. What we did not anticipate, though, is whether any of these expected differences would lead us to different substantive conclusions about student learning.

In Tables 9-13 and Figure 7, we saw that the expected pattern of parameter estimates across post-hoc methods was supported. Motivation filtering consistently produced the largest estimates of pre-test, post-test, and pre-post change in ability; leaving the data as they were observed consistently produced the lowest estimates; MI consistently produced estimates between those from the other two methods. In Table 15 however, we saw that although test statistic magnitude differed somewhat by post-hoc method, all methods suggested the same conclusions regarding the significance of the main effect of time. All three post-hoc methods’ data lead to the conclusion that, on average, examinees’ post-test NW-9 scores were statistically significantly higher than pre-test NW-9 scores.\(^{}^{19}\)

**Research question four.** Like research question three, research question four also asked whether our inferences about pre-post growth in quantitative and scientific reasoning would differ according to the method for addressing low examinee motivation. Research question four pertained to an interaction effect, though, rather than a main effect. Specifically, the fourth and final research question asked, How do the statistical significance conclusions regarding the interaction of time and number of Cluster Three courses completed at post-test on NW-9 scores compare across the three post-hoc methods for handling examinee motivation (i.e., leaving the data intact, motivation filtering, and using multiple imputation to impute plausible replacement values for rapid responses)? Based on institutional expectations, we anticipated that the more Cluster

---

\(^{19}\) I am interpreting the main effects here in light of the fact that the interaction effect was not significant. It is best practice to investigate the significance of interaction effects prior to interpreting any main effects.
Three courses completed by post-test, the more NW-9 scores would change from pre- to post-test.

In Tables 9-13, Table 15, and Figure 8, we saw that although examinees who had completed three or more Cluster Three courses by post-test grew slightly more than those who had completed less than three courses, this difference was not statistically significant, for any of the post-hoc methods.

Taken together, the results of research questions three and four suggest that although the values may have differed slightly by post-hoc method, we would have arrived at the same substantive conclusions regardless of method employed. That is, we would have concluded that examinees’ scientific and quantitative reasoning abilities increase over time, and that examinees who take more Cluster Three courses do not increase significantly more than examinees who take fewer Cluster Three courses.

There are three factors which could have produced similar results across the three post-hoc methods applied in this study. First, the data could have been MCAR. When data are MCAR, any of the post-hoc methods can be appropriately used. The results of the group comparisons between low- and adequately-motivated examinees indicated that the data were more likely MNAR, however, so I do not believe this is an adequate explanation of the results for research questions three and four.

Second, the NW-9 assessment could have been too difficult for the examinees. When assessments are too difficult for examinees, rapid responses become indistinguishable from solution-behavior responses (i.e., they all tend to be incorrect). For example, Rios et al. (2017) found that as their simulated test became more difficult it took a larger proportion of rapid responding (missingness) to produce biased means
compared to easier tests. I do not believe test difficulty is a plausible explanation for the results either, however, due to examinees’ tendency to respond correctly at rates higher than chance. At pre-test, the majority of examinees (approximately 68% of examinees) correctly responded to between 58% and 78% of the items, which is more than expected by chance alone.

A third factor that could have contributed to this outcome was the amount of missing data. Although any amount of missing data is undesirable, “trivially small” amounts are tolerable for use with traditional missing data methods (Enders, 2010, p. 55). What is considered “trivially small”, however, is somewhat unclear. In their simulation study Rios et al. (2017) found that observed means were not practically significantly different from generating means when 10% of simulees were rapidly responding—even if these 10% of simulees each rapidly responded to 50% of the items on the assessment.\textsuperscript{20} Meaningful differences were found when 25% of the sample had rapid responses, however. For this study, over 95% of examinees engaged in rapid responding for at least one item at either pre- or post-test, and 14% of examinees engaged in rapid responding for more than 10% of items at either pre- or post-test. This meant that 14% of the data were missing for the motivation filtered data, and that about 4% of the data were missing for the MI data. Given Rios et al.’s (2017) findings, it is plausible that this amount of rapid responding was not enough to witness substantive differences by post-hoc method.

**Recommendations for Researchers and Practitioners**

Upon completion of this dissertation, my recommendations to practitioners are six-fold. First, and most importantly, attempt to motivate your examinees to put forth

\textsuperscript{20} Although some means were *statistically* significantly different.
effort in providing their responses. Researchers and practitioners should continuously strive to find ways to create assessment environments in which examinees are motivated to put forth effort into their responses. For example, Wise et al. (2010) found that students were less likely to demonstrate effort when testing occurred later in the day; scheduling testing sessions in the morning may help stave off low examinee motivation. Other research has indicated the importance of proctors in fostering a motivating testing environment. Lau, Swerdzewski, Jones, Anderson, and Markle (2009) found that examinees self-reported greater effort when their testing session was conducted by proctors trained specifically to “support productive test-taking student behavior” (p. 206), for example by exuding a welcoming and respectful demeanor, and by encouraging examinees to continue to provide effort and attention to their responses throughout the testing session. Additionally, researchers have examined the effects of the following on examinee motivation: (a) notifying examinees that they will receive feedback regarding their performance, which had little to no impact on motivation (Finney et al., 2016; Swain, Williams, Hopkins, Sundre, & Finney, 2013), (b) having electronic warning messages pop-up on examinees’ computer screens when rapid responding is detected (effort-monitoring CBT), which worked well to combat low-motivation (Wise, Bhola, & Yang, 2006), and (c) shortening measures through planned missingness designs, which had a positive impact on motivation (Swain, 2015). Ultimately, post-hoc methods, which were the focus of this dissertation, should only be considered as a backup approach for situations in which examinees were not adequately motivated during the data collection phase.
My second recommendation is to measure examinee motivation (e.g., by measuring examinees’ item response times). By measuring examinee motivation you can gauge its prevalence in your data, as well as its potential to impact the inferences you make from your data. For example, by measuring examinee motivation you may find that your examinees were adequately motivated, and thus you have justification to analyze your data as they are. Gathering examinee motivation evidence can also help you better understand the experiences of your examinees, and may help justify the need to allot resources toward efforts to increase examinee motivation. Further, examinee motivation data can be provided to external users of the data (or along with the results), as is done with Programme for International Student Assessment (PISA) data (Kunter et al., 2002). Finally, if measured, examinee motivation data can be used to inform the use of post-hoc methods for addressing low motivation. Collecting *item-level* examinee motivation data may be especially advantageous in this regard because it would allow for the use of one of the modern missing data methods.

My third recommendation is for researchers and practitioners to collect data on auxiliary variables suspected to relate to the ability being measured by the assessment of interest. Such information is important for my fourth recommendation, which is to use the collected examinee motivation and auxiliary data to assess the plausibility of the MCAR assumption for motivation filtering. This might include examining relationships between missingness (i.e., whether or not an examinee would be filtered from the dataset) and other measured variables, as was done in the present study. If the MCAR assumption seems plausible, and the proportion of examinees to be filtered is not large, you may be justified in applying motivation filtering to your data.
Fifth, consider applying more than one post-hoc method and comparing the results you obtain from each. If for example you obtain essentially the same estimates from both motivation filtering and MI (which would happen if your data were MCAR), you may have justification to use either method. Comparing the results of different post-hoc methods with different assumptions about the data can help you better understand your data, and thus help you analyze and interpret it more appropriately.

Sixth and finally, I recommend that once a researcher or practitioner decides which post-hoc method to apply (or which set of results to interpret and present), that they consider and interpret their results in light of any limitations of the method they chose. For example, when applying motivation filtering, researchers and practitioners should acknowledge that their results may not represent the entire population of interest. In such instances, one might note that the results represent examinees who demonstrated adequate test-taking effort, which may tend to be examinees with moderate-to-high ability in the assessment area of interest. As another example, when applying MI, researchers and practitioners should not be tempted to make inferences about student learning at the individual student level; MI is only appropriate for group-level analyses.

**Primary Limitations of this Study**

There were four major limitations of this study. The first two limitations stemmed from the fact that the data were real, and not simulated. Thus, the true parameter values were unknown. This meant that this study was limited by the inability to definitively determine which post-hoc method best estimated the parameters of interest (the first limitation).
A second limitation, also due to the real (non-simulated) data, is that we did not know which item responses were truly associated with solution behavior and which were truly rapid guesses. That is, we did not definitively know which item responses were reflective of examinee ability. For this study, we relied on item response time information to distinguish between rapid and solution-behavior responses. However, this assumes that quick responses were guesses, and that responses associated with longer response times were obtained via solution behavior. It is quite possible that some of the responses deemed as coming from solution behavior were guesses, though. For example, a low-motivated examinee may have slowly responded to items in order to kill time during the standardized-length testing session. Or, examinees could have been distracted by their phones, or the proctors, or any number of other elements in the testing environment that could have delayed any of their item responses. Because this method only captures rapid guessing (and does not capture non-rapid guessing), some of the results we found may have alternative explanations to those described previously. For example, the differences I found between low- and adequately-motivated examinees’ “solution-behavior” responses may have been due to a misclassification of low-motivated examinees’ responses. That is, the responses I considered to have come from solution behavior may in fact have been contaminated with non-rapid guessing. So, although I attributed these group differences to differences in scientific and quantitative reasoning ability, it is possible that these group differences were due to a misclassification of responses.\(^{21}\) Not only does misclassification of responses have consequences for tests of the MCAR assumption, it also has implications for the performance of the missing data method. For

\(^{21}\) Although we note that the group differences on SAT math scores provide support for a relationship between motivation and quantitative (and scientific) reasoning ability.
example, if some of the responses in this study that were deemed as coming from solution behavior were guesses, these guesses would have been used as auxiliary variables to impute the missing values of the responses that were deemed guesses. This underscores the utility of collecting and incorporating external auxiliary information such as standardized test scores related to the ability being measured.

The third and fourth limitations of this study have to do with the MI method itself. Ideally, researchers would be able to easily adopt and apply the MI method for handling motivation-imposed missingness in their data, and subsequently easily analyze their multiply-imputed data as they would any other complete dataset. A current limitation of this method, though, is that its use with analyses in the ANOVA framework is still being developed (e.g., Finch, 2016; Grund et al., 2016; van Ginkel & Kroonenberg, 2014). Specifically, best practices for pooling the results of ANOVAs, and how to estimate appropriate standardized effect sizes, have not yet been determined. Thus, the applicability of the methods used in this study for general use are somewhat limited at present.

A final limitation of the MI method, and thus of this study, is that although MI may be the most appropriate option for handling data from low-motivated examinees, it is still not an appropriate method if the researcher or practitioner wishes to make inferences at the individual level. This limitation has been mentioned previously, but here I continue by noting that MI’s inappropriateness for individual-level inference compounds when auxiliary variables are used to impute the missing values. When auxiliary variables are used in the imputation phase, the predicted (imputed) individual-level scores explicitly contain a weighted portion of other variables. For example, in this study, examinees’
imputed values were a weighted composite of their solution-behavior responses, plus their SAT scores and Cluster Three information. Even if these weighted composite predicted scores were more accurate due to the use of the auxiliary variables, questions of fairness would arise if we were to make use of individual scores.

**Future Research**

I believe the most important future research efforts related to this study should investigate appropriate methods for pooling and summarizing ANOVA results from MI datasets. This research would serve to support the existing, but still developing, literature on this topic.

Another avenue for future research would involve simulation studies of the performance of different post-hoc methods for addressing low examinee motivation in the context of pre-post assessment. Such studies could address research questions related to the performance of the methods when the data are MNAR, or when the missingness mechanism differs from pre- to post-test, or to how well relationships between change over time and outcome variables are recovered by different methods. Other future simulation studies might aim to better study the CUMP method for distinguishing between solution-behavior and rapid responses, as it is a new method that has not been thoroughly investigated. For example, it would be beneficial to know if the CUMP method tended to produce false positives (i.e., rapid responses labeled as solution behavior) or false negatives (i.e., solution-behavior responses labeled as rapid), or how stable the CUMP thresholds are at different sample sizes, or if the CUMP thresholds are affected by item difficulty.
Other future research might investigate the impact of forcing examinees to respond to items (versus allowing them to leave items blank) on estimates of examinee ability or growth. For example, if you force examinees to respond to all items, impose missingness, and then apply MI, do your estimates of ability or growth look similar to those obtained from examinees who were allowed to skip items (and thus imposing missingness may not be necessary, but MI would still be applied)?

Another line of future research might explore the utility of not only response time and response accuracy in determining response time thresholds, but also review time. That is, for assessments that allow examinees to go back and review their responses, how do substantive results compare if thresholds are set using initial response time versus review time? What, if any, additional information can response review time provide to help us better understand examinees’ abilities?

Finally, it is important to remind readers that the best method for addressing examinee motivation is to motivate the examinees prior to and during data collection. Thus, future research investigating methods to boost examinee motivation in low-stakes testing contexts is always important.

**Summary and Conclusion**

In higher education, we collect assessment data in order to make inferences about student learning. Low examinee motivation threatens the quality of these data, and thus threatens the validity of the inferences we make about student learning based on these data. In this dissertation, the problem of low examinee test-taking effort was reframed as a missing data problem and the performance of three post-hoc methods for addressing this problem were compared.
The results showed that motivation filtering (listwise deletion) produced the highest estimates of examinee ability and growth over time. This may have been because the data were MNAR; that is, this may have been because examinees who were filtered out of the dataset may have been lower-ability examinees. Leaving the data as they were observed, replete with rapid responses, produced the lowest estimates of examinee ability and growth over time. This may have been because the observed, rapid, responses were more likely to be incorrect than if the responses had been provided using solution behavior. That is, this may have been because examinees’ rapid responses had a chance level of being correct, yet the examinees’ abilities were above chance levels (even though the examinees who rapidly responded may have been of lower ability). Finally, MI at the item level with solution-behavior items as auxiliary variables produced estimates of examinee ability and growth over time between the estimates from the previous two methods. This may have been because MI was able to use the auxiliary variables to transition the data from MNAR to MAR, rendering MI appropriate for producing unbiased estimates. Because this study was applied in nature, however, these conjectures could not be confirmed.

Ultimately, it is recommended that practitioners and researchers use these results to, at a minimum, acknowledge that the magnitude of their estimated outcomes may depend on the method they use to address low examinee motivation. Further, I encourage practitioners and researchers to consider and assess the appropriateness of various post-hoc methods for addressing low-examinee motivation in their low-stakes testing data.
Appendix A: Obtaining parameter estimates and multiparameter significance tests in SAS

Generally, when applying mixed ANOVA models practitioners are interested in obtaining tests of the model effects (i.e., main effects and interactions). For this study, however, we were also interested in the parameter estimates (and standard errors) that produced the model effects (i.e., the seven effect-coded parameters). We wanted to compare these estimates across the three post-hoc methods. To obtain both parameter estimates and tests of model effects for each of the three post-hoc methods, multiple steps were performed.

To obtain parameter estimates of the multiple regression model where NW-9 scores were predicted by time (effect coded), three effect-coded variables for the number of Cluster Three courses completed by post-test, and three effect-coded interaction terms, the PROC MIXED procedure was used. PROC MIXED, which employs maximum likelihood estimation, was used instead of PROC GLM because PROC GLM does not provide these parameter estimates.

To obtain statistical significance tests of the two main effects and the interaction, multiparameter tests were performed. For the unaltered and motivation filtered data, multivariate Wald tests were performed. Essentially, these tests evaluate a group of parameters in comparison to a null value. For example, for the main effect of number of courses the null hypothesis is that the set of the three effect-coded number of courses variables are equal to zero. The equation for the multivariate Wald test statistic is as follows (as provided by Enders, 2010):

\[
\omega = (\hat{\theta} - \theta_0)^T \text{var}(\hat{\theta})^{-1}(\hat{\theta} - \theta_0)
\]

Where \(\hat{\theta}\) is the vector of parameter estimates to be tested (as a group), \(\theta_0\) is the vector of null values (e.g., zeros), and \(\text{var}(\hat{\theta})\) is the covariance matrix for the parameters being tested (and only those parameters). The resulting test statistic, \(\omega\), follows a chi-square distribution with \(df_{\text{effect}}\) equal to the number of parameters being tested (and has an associated \(df_{\text{error}}\) equal to \((N - a)(b - 1)\), where \(N\) is the number of cases, \(a\) is the number of between-subjects groups, and \(b\) is the number of within-subjects groups).

Although the multivariate Wald test can be computed manually, the same results are achieved by using the PROC GLM procedure. For example, for this study the PROC GLM procedure was performed with time, number of courses, and their interaction predicting NW-9 scores (variables were not effect coded for this procedure). The resulting \(F\)-tests for each model effect matched those obtained by computing the multivariate Wald test manually and transforming the chi-square test statistic to an \(F\)-statistic. Thus, for this study, both PROC MIXED and PROC GLM were performed for the unaltered and motivation filtered data.

To obtain statistical significance tests of the model effects for the MI data, the \(D_1\) multiparameter test was performed. The \(D_1\) test is analogous to the multivariate Wald test, but is more appropriate for MI data because it accounts for within- and between-
imputation variance information. The equation for the $D_1$ test statistic is as follows (as provided by Enders, 2010):

$$D_1 = \frac{1}{k} (\bar{\theta} - \theta_0)^T (\bar{V}^{-1})(\bar{\theta} - \theta_0)$$

(8)

Where $k$ is the number of parameters being tested, $\bar{\theta}$ is the vector of parameters being tested, $\theta_0$ is the vector of null parameter values, and $\bar{V}_T$ is the total parameter covariance matrix, given by the following equation (from Enders, 2010):

$$\bar{V}_T = \left[ 1 + \left[ \frac{(1 + m^{-1}) tr(V_B V_W^{-1})}{k} \right] \right] V_W$$

(9)

Where $V_B$ and $V_W$ are the between- and within-imputation covariance matrices for the $k$ parameters.

In SAS, the $D_1$ multiparameter test can be performed within the PROC MIANALYZE procedure, which provides an $F$-statistic. The $df_{effect}$ for the $D_1$ test are the same as those for the multivariate Wald test (i.e., the number of parameters being tested). The $df_{error}$ are quite different however, because they take into account the number of $m$ imputations performed. When $(km - k) > 4$, the $df_{error}$ for the $D_1$ statistic are as follows (from Enders, 2010):

$$df_{error} = 4 + (km - k - 4) \left[ 1 + \left[ 1 - \frac{2}{km - k} \right] \frac{1}{(1 + m^{-1}) tr(V_B V_W^{-1})} \right]^2$$

(10)

When $(km - k) < 4$, the $df_{error}$ for the $D_1$ statistic are as follows (from Enders, 2010):

$$df_{error} = \frac{(km - k)(1 + \frac{1}{k})}{2} \left[ 1 + \frac{1}{(1 + m^{-1}) tr(V_B V_W^{-1})} \right]^2$$

(11)

See Appendix D for SAS syntax.
Appendix B: R syntax for setting CUMP thresholds

# K. Foelber
# Spring 2017
# setting item response time thresholds using the CUMP method (Guo et al., 2016)
# using pre-post NW9 data

#------------------------------------------------------------------------------------------
#------------------------------------------------------------------------------------------
# this syntax requires two files:
# 1. "ALL.csv" = item scores & item response times by student ID, for both fall and spring
# 2. "nw9_numMCoptions.csv" = number of response options by item

setwd("N:/AA/CARS/CARS-Common/Graduate Students/GAFILES/Kelly/diss")
install.packages("psych")
require(psych)

nw9 <- read.csv(file="ALL.csv", header=TRUE, sep="",")
names(nw9)[names(nw9)=="...ID"] <- "ID"

nw9resp <- read.csv(file="nw9_numMCoptions.csv", header=TRUE, sep="",")
names(nw9resp)[names(nw9resp)=="X..options"] <- "options"

# ------------------------

fall.results <- data.frame(FA_threshold=NA)
for(item in 1:66) {
  scorecol <- paste("FAsc",item,sep="")
scores <- nw9[, scorecol]
timecol <- paste("FAt",item,sep="")
time <- nw9[, timecol]

  freq <- as.data.frame.matrix(table(time, scores))
colnames(freq) <- c("incorrect","correct")
freq$srt <- as.numeric(rownames(freq))
freq$cum <- cumsum(freq$correct)
freq$cumprop <- freq$cum/388
freq$freq <- freq$incorrect + freq$correct
freq$cumfreq <- cumsum(freq$freq)
freq$relcumprop <- freq$cum/freq$cumfreq
freq$chance <- 1/(nw9resp[item,2])}
for (i in 1:nrow(freq)) {freq$above[i] <- ifelse(freq$relcumprop[i] >= freq$chance[i], 1, 0)}
for (i in 1:nrow(freq)) {freq$threshold.rt[i] <- ifelse(freq$above[i]==0, freq$rt[i+1], 0)}
fell.threshold <- max(freq$threshold.rt)

fall.results[item,1] <- fall.threshold
}

# ----------------------------- #

spring.results <- data.frame(SP_threshold=NA)

for(item in 1:66) {
  scorecol <- paste("SPsc",item,sep="")
  scores <- nw9[, scorecol]
  timecol <- paste("SPt",item,sep="")
  time <- nw9[, timecol]

  freq <- as.data.frame.matrix(table(time, scores))
  colnames(freq) <- c("incorrect","correct")
  freq$rt <- as.numeric(rownames(freq))
  freq$cum <- cumsum(freq$correct)
  freq$cumprop <- freq$cum/388
  freq$freq <- freq$incorrect + freq$correct
  freq$cumfreq <- cumsum(freq$freq)
  freq$relcumprop <- freq$cum/freq$cumfreq
  freq$chance <- 1/(nw9resp[item,2])
  for (i in 1:nrow(freq)) {freq$above[i] <- ifelse(freq$relcumprop[i] >= freq$chance[i], 1, 0)}
  for (i in 1:nrow(freq)) {freq$threshold.rt[i] <- ifelse(freq$above[i]==0, freq$rt[i+1], 0)}
spring.threshold <- max(freq$threshold.rt)

  spring.results[item,1] <- spring.threshold
}

# ----------------------------- #

final.results <- cbind(fall.results,spring.results)
final.results$item <- as.numeric(rownames(final.results))
final.results
write.table(final.results, "CUMP_finalresults.csv", row.names=FALSE, sep="","")
Appendix C: Blimp syntax for imputing data

DATA: N:\AA\CARS\CARS-Common\Graduate
Students\GAFILES\Kelly\diss\forBLIMP.dat;

VARIABLES: VAR1 VAR2 VAR3 VAR4 VAR5 VAR6 VAR7 VAR8 VAR9 VAR10
VAR11 VAR12 VAR13 VAR14 VAR15 VAR16 VAR17 VAR18 VAR19 VAR20
VAR21 VAR22 VAR23 VAR24 VAR25 VAR26 VAR27 VAR28 VAR29 VAR30
VAR31 VAR32 VAR33 VAR34 VAR35 VAR36 VAR37 VAR38 VAR39 VAR40
VAR41 VAR42 VAR43 VAR44 VAR45 VAR46 VAR47 VAR48 VAR49 VAR50
VAR51 VAR52 VAR53 VAR54 VAR55 VAR56 VAR57 VAR58 VAR59 VAR60
VAR61 VAR62 VAR63 VAR64 VAR65 VAR66 VAR67 VAR68 VAR69 VAR70
VAR71 VAR72 VAR73 VAR74 VAR75 VAR76 VAR77 VAR78 VAR79 VAR80
VAR81 VAR82 VAR83 VAR84 VAR85 VAR86 VAR87 VAR88 VAR89 VAR90
VAR91 VAR92 VAR93 VAR94 VAR95 VAR96 VAR97 VAR98 VAR99 VAR100
VAR101 VAR102 VAR103 VAR104 VAR105 VAR106 VAR107 VAR108 VAR109
VAR110 VAR111 VAR112 VAR113 VAR114 VAR115 VAR116 VAR117 VAR118
VAR119 VAR120 VAR121 VAR122 VAR123 VAR124 VAR125 VAR126 VAR127
VAR128 VAR129 VAR130 VAR131 VAR132 VAR133 VAR134 VAR135 VAR136;

ORDINAL: VAR1 VAR2 VAR3 VAR4 VAR5 VAR6 VAR7 VAR8 VAR9 VAR10
VAR11 VAR12 VAR13 VAR14 VAR15 VAR16 VAR17 VAR18 VAR19 VAR20
VAR21 VAR22 VAR23 VAR24 VAR25 VAR26 VAR27 VAR28 VAR29 VAR30
VAR31 VAR32 VAR33 VAR34 VAR35 VAR36 VAR37 VAR38 VAR39 VAR40
VAR41 VAR42 VAR43 VAR44 VAR45 VAR46 VAR47 VAR48 VAR49 VAR50
VAR51 VAR52 VAR53 VAR54 VAR55 VAR56 VAR57 VAR58 VAR59 VAR60
VAR61 VAR62 VAR63 VAR64 VAR65 VAR66 VAR67 VAR68 VAR69 VAR70
VAR71 VAR72 VAR73 VAR74 VAR75 VAR76 VAR77 VAR78 VAR79 VAR80
VAR81 VAR82 VAR83 VAR84 VAR85 VAR86 VAR87 VAR88 VAR89 VAR90
VAR91 VAR92 VAR93 VAR94 VAR95 VAR96 VAR97 VAR98 VAR99 VAR100
VAR101 VAR102 VAR103 VAR104 VAR105 VAR106 VAR107 VAR108 VAR109
VAR110 VAR111 VAR112 VAR113 VAR114 VAR115 VAR116 VAR117 VAR118
VAR119 VAR120 VAR121 VAR122 VAR123 VAR124 VAR125 VAR126 VAR127
VAR128 VAR129 VAR130 VAR131 VAR132 VAR133 VAR134 VAR135 VAR136;

NOMINAL: ;

MISSING: -99;

MODEL: ~ VAR1 VAR2 VAR3 VAR4 VAR5 VAR6 VAR7 VAR8 VAR9 VAR10
VAR11 VAR12 VAR13 VAR14 VAR15 VAR16 VAR17 VAR18 VAR19 VAR20
VAR21 VAR22 VAR23 VAR24 VAR25 VAR26 VAR27 VAR28 VAR29 VAR30
VAR31 VAR32 VAR33 VAR34 VAR35 VAR36 VAR37 VAR38 VAR39 VAR40
VAR41 VAR42 VAR43 VAR44 VAR45 VAR46 VAR47 VAR48 VAR49 VAR50
VAR51 VAR52 VAR53 VAR54 VAR55 VAR56 VAR57 VAR58 VAR59 VAR60
VAR61 VAR62 VAR63 VAR64 VAR65 VAR66 VAR67 VAR68 VAR69 VAR70
VAR71 VAR72 VAR73 VAR74 VAR75 VAR76 VAR77 VAR78 VAR79 VAR80
VAR81 VAR82 VAR83 VAR84 VAR85 VAR86 VAR87 VAR88 VAR89 VAR90
VAR91 VAR92 VAR93 VAR94 VAR95 VAR96 VAR97 VAR98 VAR99 VAR100
VAR101 VAR102 VAR103 VAR104 VAR105 VAR106 VAR107 VAR108 VAR109
VAR110 VAR111 VAR112 VAR113 VAR114 VAR115 VAR116 VAR117 VAR118
VAR119 VAR120 VAR121 VAR122 VAR123 VAR124 VAR125 VAR126 VAR127
VAR128 VAR129 VAR130 VAR131 VAR132 VAR133 VAR134 VAR135 VAR136;

NIMPS: 100;

THIN: 10000;

BURN: 10000;

SEED: 14;

OUTFILE: N:\AA\CARS\CARS-Common\Graduate
Students\GAFILES\Kelly\diss\BLIMPut2_kjf.csv;

OPTIONS: stacked nopsr csv clmean prior2 hov;
Appendix D: SAS syntax

* K. Foelber;
* dissertation analyses;
* Spring 2017;

* UNALTERED DATA

proc import out=donnothing
datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAPFILES\Kelly\diss\do_nothing_long.sav"
replace;
run;
proc import out=donnothing_wide
datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAPFILES\Kelly\diss\do_nothing.sav"
replace;
run;

data donnothing_wide; set donnothing_wide;
fa_tot=sum(of fasc1-fasc66);
sp_tot=sum(of spsc1-spsc66);
drop fasc1-fasc66 spsc1-spsc66;
diff_tot=sp_tot-fa_tot;
run;

*code num3courses into 4 groups;
data donnothing; set donnothing;
if num3courses=0 then num3courses_4=0;
if num3courses=1 then num3courses_4=1;
if num3courses=2 then num3courses_4=2;
if num3courses GE 3 then num3courses_4=3;
run;
data donnothing_wide; set donnothing_wide;
if num3courses=0 then num3courses_4=0;
if num3courses=1 then num3courses_4=1;
if num3courses=2 then num3courses_4=2;
if num3courses GE 3 then num3courses_4=3;
run;

*checking out data;
proc sort data=donnothing_wide; by numc3courses_4; run;
proc corr data=donnothing_wide cov;
var fa_tot sp_tot;
by numc3courses_4;
run;
proc means data=donnothing_wide n mean stderr std skew kurt;
var fa_tot sp_tot diff_tot;
by numc3courses_4;
run;
proc sort data=donnothing; by numc3courses_4; run;
proc means data=donnothing n mean std;
var score;
by numc3courses_4;
run;

*creating effect-coded IVs (time and numc3courses);
data donothing; set donothing;
if time_dummy=0 then time_e=-1;
if time_dummy=1 then time_e=1;
if numc3courses_4=0 then numc3_e1=-1;
if numc3courses_4=1 then numc3_e1=1;
if numc3courses_4=2 then numc3_e1=0;
if numc3courses_4=3 then numc3_e1=0;

if numc3courses_4=0 then numc3_e2=-1;
if numc3courses_4=1 then numc3_e2=0;
if numc3courses_4=2 then numc3_e2=1;
if numc3courses_4=3 then numc3_e2=0;

if numc3courses_4=0 then numc3_e3=-1;
if numc3courses_4=1 then numc3_e3=0;
if numc3courses_4=2 then numc3_e3=0;
if numc3courses_4=3 then numc3_e3=1;
run;

*running split plot anova with effect coding;
proc mixed data=donothing;
class time_dummy;
model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1
   time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
repeated time_dummy/subject=ID type=cs r rcorr;
ods output SolutionF=paraDN Tests3=FtestsDN covb=covDN;
run;

*running proc mixed again with CS not assumed - UNSTRUCTURED, free by group;
proc mixed data=donothing;
class time_dummy numc3courses_4;
model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1
   time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
repeated time_dummy/subject=ID group=numc3courses_4 type=un r rcorr;
run;

*running split plot ANOVA through proc glm to get sig tests;
proc glm data=donothing_wide;
class numc3courses_4;
model fa_tot sp_tot=numc3courses_4 /solution nouni;
repeated time 2 (0 1)/printe;
run;
quit;
*import listwise data;
proc import out=listwise
datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAFILES\Kelly\diss\listwise_long.sav" replace;
run;
proc import out=listwise_wide
datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAFILES\Kelly\diss\listwise.sav" replace;
run;
data listwise_wide; set listwise;
fa_tot=sum(of fasc1-fasc66);
sp_tot=sum(of spsci-spsc66);
drop fasc1-fasc66 spsci-spsc66;
diff_tot=sp_tot-fa_tot;
run;
*code numc3courses into 4 groups;
data listwise; set listwise;
if numc3courses=0 then numc3courses_4=0;
if numc3courses=1 then numc3courses_4=1;
if numc3courses=2 then numc3courses_4=2;
if numc3courses GE 3 then numc3courses_4=3;
run;
data listwise_wide; set listwise;
if numc3courses=0 then numc3courses_4=0;
if numc3courses=1 then numc3courses_4=1;
if numc3courses=2 then numc3courses_4=2;
if numc3courses GE 3 then numc3courses_4=3;
run;
*creating effect coded IVs (time and numc3courses);
data listwise; set listwise;
if time_dummy=0 then time_e=-1;
if time_dummy=1 then time_e=1;
if numc3courses_4=0 then numc3_e1=-1;
if numc3courses_4=1 then numc3_e1=1;
if numc3courses_4=2 then numc3_e1=0;
if numc3courses_4=3 then numc3_e1=0;
if numc3courses_4=0 then numc3_e2=-1;
if numc3courses_4=1 then numc3_e2=0;
if numc3courses_4=2 then numc3_e2=1;
if numc3courses_4=3 then numc3_e2=0;
if numc3courses_4=0 then numc3_e3=-1;
if numc3courses_4=1 then numc3_e3=0;
if numc3courses_4=2 then numc3_e3=0;
if numc3courses_4=3 then numc3_e3=1;
run;
* checking out data;
  proc sort data=listwise_wide; by numc3courses_4; run;
  proc corr data=listwise_wide cov;
    var fa_tot sp_tot;
    by numc3courses_4;
  run;
  proc means data=listwise_wide n mean std skew kurt;
    var fa_tot sp_tot diff_tot;
    by numc3courses_4;
  run;
  proc sort data=listwise; by numc3courses_4; run;
  proc means data=listwise n mean std;
    var score;
    by numc3courses_4;
  run;

*running split plot anova with effect coding;
  proc mixed data=listwise;
    class time_dummy;
    model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1
              time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
    repeated time_dummy/subject=ID type=cs r rcorr;
    ods output SolutionF=paraLW Tests3=FtestsLW covb=covLW;
  run;

*running proc mixed again with CS not assumed - UNSTRUCTURED, free by group;
  proc mixed data=listwise;
    class time_dummy numc3courses_4;
    model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1
              time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
    repeated time_dummy/subject=ID group=numc3courses_4 type=un r rcorr;
  run;

*running split plot ANOVA through proc glm to get sig tests;
  proc glm data=listwise_wide;
    class numc3courses_4;
    model fa_tot sp_tot=numc3courses_4 /solution nouni;
    repeated time 2 (0 1)/printe;
  run;
  quit;
* reading in Blimp MI data;

```sas
data miout;
infile "N:\AA\CARS\CARS-Common\Graduate Students\GAPFILES\Kelly\diss\BLIMPout2 kjf dontwriteoverme.csv" dlm=',';
input _Imputation_ fasbsc1-fasbsc66 spsbsc1-spsbsc66 sat1math sat1verb numc3courses c3GPA;
fa_tot=sum(of fasbsc1-fasbsc66);
sp_tot=sum(of spsbsc1-spsbsc66);
rand_id = _N_;
drop fasbsc1-fasbsc66 spsbsc1-spsbsc66;
run;
```

```sas
proc means data=miout; var fa_tot sp_tot sat1math sat1verb numc3courses c3GPA;
class _Imputation_;
run;
```

```sas
proc sort data=miout by rand_id _imputation_ sat1math sat1verb numc3courses c3GPA fa_tot sp_tot;
run;
```

```sas
proc transpose data=miout out=miout_tot_t; var fa_tot sp_tot; by rand_id _imputation_ sat1math sat1verb numc3courses c3GPA;
run;
```

```sas
proc sort data=miout_tot_t by _imputation_; run;
data miout_tot_t; set miout_tot_t;
if _NAME_="fa_tot" then time_dummy=0;
if _NAME_="sp_tot" then time_dummy=1;
score=COL1;
run;
```

```sas
data miout_tot_t; set miout_tot_t;
if numc3courses=0 then numc3courses_4=0;
if numc3courses=1 then numc3courses_4=1;
if numc3courses=2 then numc3courses_4=2;
if numc3courses GE 3 then numc3courses_4=3;
run;
```

*run descriptives on each dataset (then pool);

```sas
data miout_tot; set miout; diff_tot=sp_tot-fa_tot; run;
```

```sas
proc sort data=miout_tot by _imputation_; run;
proc univariate data=miout_tot noprint;
var fa_tot sp_tot diff_tot;
output out=outuni mean=fa_tot sp_tot diff_tot stderr=se_fa_tot se_sp_tot se_diff_tot std=std_fa_tot std_sp_tot std_diff_tot;
by _imputation_;
run;
```

```sas
data miout_tot; set miout_tot;
if numc3courses=0 then numc3courses_4=0;
```

if numc3courses=1 then numc3courses_4=1;
if numc3courses=2 then numc3courses_4=2;
if numc3courses GE 3 then numc3courses_4=3;
run;

proc sort data=miout_tot; by _imputation_ numc3courses_4; run;
proc univariate data=miout_tot noprint;
var fa_tot sp_tot;
output out=outuniz mean=fa_tot sp_tot stderr=se_fa_tot se_sp_tot std=std_fa_tot std_sp_tot;
by _imputation_ numc3courses_4;
run;

*pooling descriptives;
proc mianalyze data=outuni edf=380;
 modeleffects fa_tot sp_tot diff_tot std_fa_tot std_sp_tot;
  stderr se_fa_tot se_sp_tot se_diff_tot std_diff_tot;
run;
proc sort data=outuniz; by numc3courses_4 _imputation_; run;
proc mianalyze data=outuniz edf=380;
  by numc3courses_4;
  modeleffects fa_tot sp_tot std_fa_tot std_sp_tot;
  stderr se_fa_tot se_sp_tot std_diff_tot;
run;

*creating effect coded IVs (time and numc3courses);
data miout_tot_t; set miout_tot_t;
  if time_dummy=0 then time_e=-1;
  if time_dummy=1 then time_e=1;
  if numc3courses_4=0 then numc3_e1=-1;
  if numc3courses_4=1 then numc3_e1=1;
  if numc3courses_4=2 then numc3_e2=0;
  if numc3courses_4=3 then numc3_e2=1;
  if numc3courses_4=0 then numc3_e3=-1;
  if numc3courses_4=1 then numc3_e3=0;
  if numc3courses_4=2 then numc3_e3=1;
  if numc3courses_4=3 then numc3_e3=0;
run;

* run mixed design/split plot anova with effect coding on each dataset;
proc mixed data=miout_tot_t;
  class time_dummy;
  model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1 time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
repeated time_dummy/subject=rand_id type=cs rrcorr;
by _Imputation_
ods output SolutionF=paraMI Tests3=FtestsMI covb=covMI
FitStatistics=fitMI;
run;

*running proc mixed again with CS not assumed-- unstructured + free by group;
proc mixed data=miout_tot_t;
class time_dummy numc3courses_4;
model score=time_e numc3_e1 numc3_e2 numc3_e3 time_e*numc3_e1
time_e*numc3_e2 time_e*numc3_e3/solution covb chisq;
repeated time_dummy/subject=rand_id group=numc3courses_4 type=un r
rcorr;
by _Imputation_
ods output FitStatistics=fitMI_UN;
run;

data fitMI_UN; set fitMI_UN; descr_UN=descr; value_UN=value; drop descr
value; run;

data fit; merge fitMI fitMI_UN; by _imputation_; run;

data fit; set fit;
if descr=descr_UN then checkname=1;
if value LT value_UN then csbetter=1;
if descr="-2 Res Log Likelihood" then delete;
run;

proc freq data=fit; tables checkname csbetter; run;

* re-naming the interaction effects-- mianalyze doesn't like the asterisk;
data paraMI; set paraMI;
if effect="time_e*numc3_e1" then effect="int_time_e1";
if effect="time_e*numc3_e2" then effect="int_time_e2";
if effect="time_e*numc3_e3" then effect="int_time_e3";
run;
data covMI; set covMI;
if effect="time_e*numc3_e1" then effect="int_time_e1";
if effect="time_e*numc3_e2" then effect="int_time_e2";
if effect="time_e*numc3_e3" then effect="int_time_e3";
run;

* pooling parms and SEs across 100 datasets;
proc mianalyze parms=paraMI covb(effectvar=rowcol)=covMI edf=380;
modeleffects Intercept time_e numc3_el numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
stderr __ Intercept time_e numc3_el numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
time: test time_e=0 /mult; *don't need the mult here, but if I put it
I get F instead of t statistic;
numc3: test numc3_e1=numc3_e2=numc3_e3=0 /mult;
int: test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
* pool results across first 3, 5, 10, 25, and 50 datasets;
* take subsamples of 100 MI datasets, save as new datasets;
* run proc mianalyze on each of these datasets;

*descriptives: take subsamples;

```
data outuni3; set outuni; if _imputation_ GT 3 then delete; run;
data outuni5; set outuni; if _imputation_ GT 5 then delete; run;
data outuni10; set outuni; if _imputation_ GT 10 then delete; run;
data outuni25; set outuni; if _imputation_ GT 25 then delete; run;
data outuni50; set outuni; if _imputation_ GT 50 then delete; run;
```

*pooling descriptives;

```
proc mianalyze data=outuni3 edf=380; modeleffects fa_tot sp_tot
diff_tot; stderr se_fa_tot se_sp_tot se_diff_tot; run;
proc mianalyze data=outuni5 edf=380; modeleffects fa_tot sp_tot
diff_tot; stderr se_fa_tot se_sp_tot se_diff_tot; run;
proc mianalyze data=outuni10 edf=380; modeleffects fa_tot sp_tot
diff_tot; stderr se_fa_tot se_sp_tot se_diff_tot; run;
proc mianalyze data=outuni25 edf=380; modeleffects fa_tot sp_tot
diff_tot; stderr se_fa_tot se_sp_tot se_diff_tot; run;
proc mianalyze data=outuni50 edf=380; modeleffects fa_tot sp_tot
diff_tot; stderr se_fa_tot se_sp_tot se_diff_tot; run;
```

* split plot parms + SEs: take subsamples;

```
data parami3; set parami; if _imputation_ GT 3 then delete; run;
data parami5; set parami; if _imputation_ GT 5 then delete; run;
data parami10; set parami; if _imputation_ GT 10 then delete; run;
data parami25; set parami; if _imputation_ GT 25 then delete; run;
data parami50; set parami; if _imputation_ GT 50 then delete; run;
```

```
data covMI3; set covMI; if _imputation_ GT 3 then delete; run;
data covMI5; set covMI; if _imputation_ GT 5 then delete; run;
data covMI10; set covMI; if _imputation_ GT 10 then delete; run;
data covMI25; set covMI; if _imputation_ GT 25 then delete; run;
data covMI50; set covMI; if _imputation_ GT 50 then delete; run;
```

*pooling split plot parms + SEs by subsample;

```
proc mianalyze parms=paraMI3 covb(effectvar=rowcol)=covMI3 edf=380;
   modeleffects Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
   int_time_e2 int_time_e3;
   stderr Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
   int_time_e2 int_time_e3;
   time: test time_e=0 /mult;
   numc3: test numc3_e1=numc3_e2=numc3_e3=0 /mult;
   int: test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
proc mianalyze parms=paraMI5 covb(effectvar=rowcol)=covMI5 edf=380;
   modeleffects Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
   int_time_e2 int_time_e3;
```
stderr    Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
  time:  test time_e=0 /mult;
numc3:  test numc3_e1=numc3_e2=numc3_e3=0 /mult;
  int:   test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
proc mianalyze parms=paraMI10 covb(effectvar=rowcol)=covMI10 edf=380;
  modeleffects Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
stderr    Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
  time:  test time_e=0 /mult;
numc3:  test numc3_e1=numc3_e2=numc3_e3=0 /mult;
  int:   test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
proc mianalyze parms=paraMI25 covb(effectvar=rowcol)=covMI25 edf=380;
  modeleffects Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
stderr    Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
  time:  test time_e=0 /mult;
numc3:  test numc3_e1=numc3_e2=numc3_e3=0 /mult;
  int:   test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
proc mianalyze parms=paraMI50 covb(effectvar=rowcol)=covMI50 edf=380;
  modeleffects Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
stderr    Intercept time_e numc3_e1 numc3_e2 numc3_e3 int_time_e1
int_time_e2 int_time_e3;
  time:  test time_e=0 /mult;
numc3:  test numc3_e1=numc3_e2=numc3_e3=0 /mult;
  int:   test int_time_e1=int_time_e2=int_time_e3=0 /mult;
run;
References


Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can


relationship between time of testing and test-taking effort. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.


Table 1

*Response Accuracy by Response Time for Item 8 of the NW-9 Pre-Test Administration*

<table>
<thead>
<tr>
<th>Response time (seconds)</th>
<th>Number of examinees who responded incorrectly</th>
<th>Number of examinees who responded correctly</th>
<th>Cumulative number of examinees who responded correctly</th>
<th>Cumulative number of examinees</th>
<th>Cumulative proportion of examinees who responded correctly</th>
<th>Probability of correct response by chance alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0.0000</td>
<td>0.3333</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0.0000</td>
<td>0.3333</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0.0000</td>
<td>0.3333</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0.1111</td>
<td>0.3333</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>0.2000</td>
<td>0.3333</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>0.1818</td>
<td>0.3333</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>0.1667</td>
<td>0.3333</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>14</td>
<td>0.2143</td>
<td>0.3333</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>0.3125</td>
<td>0.3333</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>19</td>
<td>0.3158</td>
<td>0.3333</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>25</td>
<td>0.4400</td>
<td><strong>0.3333</strong></td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>28</td>
<td>0.4643</td>
<td>0.3333</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>29</td>
<td>0.4828</td>
<td>0.3333</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>126</td>
<td>0</td>
<td>1</td>
<td>308</td>
<td>386</td>
<td>0.7979</td>
<td>0.3333</td>
</tr>
<tr>
<td>128</td>
<td>0</td>
<td>1</td>
<td>309</td>
<td>387</td>
<td>0.7985</td>
<td>0.3333</td>
</tr>
<tr>
<td>129</td>
<td>1</td>
<td>0</td>
<td>309</td>
<td>388</td>
<td>0.7964</td>
<td>0.3333</td>
</tr>
</tbody>
</table>

*Note. N=388*
Table 2

*Example Dataset*

<table>
<thead>
<tr>
<th>$Y_{obs}$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>--</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>--</td>
<td>1</td>
</tr>
<tr>
<td>--</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3

*Example Dataset for Deletion*

<table>
<thead>
<tr>
<th>Case</th>
<th>Y₁</th>
<th>Y₂</th>
<th>Y₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>--</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>--</td>
<td>9</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>--</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>
Table 4

**Sample Descriptive Information**

<table>
<thead>
<tr>
<th># of students (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohort</strong></td>
<td></td>
</tr>
<tr>
<td>FA08-SP10</td>
<td>159 (41%)</td>
</tr>
<tr>
<td>FA10-SP12</td>
<td>99 (26%)</td>
</tr>
<tr>
<td>FA11-SP13</td>
<td>130 (34%)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>244 (63%)</td>
</tr>
<tr>
<td>Male</td>
<td>144 (37%)</td>
</tr>
<tr>
<td><strong>Ethnic Group</strong></td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>Asian</td>
<td>20 (5%)</td>
</tr>
<tr>
<td>Black</td>
<td>10 (3%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>15 (4%)</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>50 (13%)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>3 (1%)</td>
</tr>
<tr>
<td>White</td>
<td>270 (70%)</td>
</tr>
<tr>
<td>Not Specified</td>
<td>18 (5%)</td>
</tr>
<tr>
<td><strong>Virginia Residency</strong></td>
<td></td>
</tr>
<tr>
<td>In-state resident</td>
<td>268 (69%)</td>
</tr>
<tr>
<td>Out-of-state resident</td>
<td>120 (31%)</td>
</tr>
<tr>
<td><strong>Country of Origin</strong></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>379 (98%)</td>
</tr>
<tr>
<td>Non-USA country</td>
<td>9 (2%)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Rounded to the nearest whole number
Table 5 (continued on next page)

*Thresholds and Threshold-Setting Methods for Each Item at Each Time Point*

<table>
<thead>
<tr>
<th>Item</th>
<th>Pre-test method</th>
<th>Threshold (seconds)</th>
<th>Post-test method</th>
<th>Threshold (seconds)</th>
<th>Difference in thresholds (post-pre)</th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VI</td>
<td>10.5</td>
<td>VI</td>
<td>8.0</td>
<td>-2.5</td>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>VI</td>
<td>12.5</td>
<td>CUMP</td>
<td>30.0</td>
<td>17.5</td>
<td>0.23</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>CUMP</td>
<td>2.0</td>
<td>VI</td>
<td>8.0</td>
<td>6.0</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>4</td>
<td>VI</td>
<td>20.0</td>
<td>VI</td>
<td>12.0</td>
<td>-8.0</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>CUMP</td>
<td>20.0</td>
<td>CUMP</td>
<td>16.0</td>
<td>-4.0</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>CUMP</td>
<td>20.0</td>
<td>VI</td>
<td>8.5</td>
<td>-11.5</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>7</td>
<td>VI</td>
<td>10.0</td>
<td>CUMP</td>
<td>12.0</td>
<td>2.0</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>8</td>
<td>CUMP</td>
<td>22.0</td>
<td>CUMP</td>
<td>17.0</td>
<td>-5.0</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>9</td>
<td>CUMP</td>
<td>7.0</td>
<td>CUMP</td>
<td>6.0</td>
<td>1.0</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>10</td>
<td>CUMP</td>
<td>37.0</td>
<td>CUMP</td>
<td>43.0</td>
<td>6.0</td>
<td>0.61</td>
<td>0.65</td>
</tr>
<tr>
<td>11</td>
<td>CUMP</td>
<td>7.0</td>
<td>CUMP</td>
<td>2.0</td>
<td>5.0</td>
<td>0.49</td>
<td>0.68</td>
</tr>
<tr>
<td>12</td>
<td>CUMP</td>
<td>4.0</td>
<td>CUMP</td>
<td>3.0</td>
<td>-1.0</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>13</td>
<td>CUMP</td>
<td>3.0</td>
<td>VI</td>
<td>3.5</td>
<td>0.5</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>14</td>
<td>VI</td>
<td>11.0</td>
<td>VI</td>
<td>9.5</td>
<td>-1.5</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>15</td>
<td>VI</td>
<td>26.5</td>
<td>VI</td>
<td>24.0</td>
<td>-2.5</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>16</td>
<td>VI</td>
<td>4.5</td>
<td>CUMP</td>
<td>2.0</td>
<td>-2.5</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>17</td>
<td>VI</td>
<td>7.0</td>
<td>VI</td>
<td>5.0</td>
<td>-2.0</td>
<td>0.73</td>
<td>0.86</td>
</tr>
<tr>
<td>18</td>
<td>CUMP</td>
<td>3.0</td>
<td>CUMP</td>
<td>2.0</td>
<td>1.0</td>
<td>0.52</td>
<td>0.57</td>
</tr>
<tr>
<td>19</td>
<td>CUMP</td>
<td>10.0</td>
<td>CUMP</td>
<td>2.0</td>
<td>-8.0</td>
<td>0.40</td>
<td>0.58</td>
</tr>
<tr>
<td>20</td>
<td>VI</td>
<td>4.5</td>
<td>CUMP</td>
<td>5.0</td>
<td>0.5</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>21</td>
<td>VI</td>
<td>9.0</td>
<td>CUMP</td>
<td>5.0</td>
<td>4.0</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>22</td>
<td>CUMP</td>
<td>23.0</td>
<td>CUMP</td>
<td>14.0</td>
<td>-9.0</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>23</td>
<td>CUMP</td>
<td>3.0</td>
<td>VI</td>
<td>5.0</td>
<td>2.0</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>24</td>
<td>CUMP</td>
<td>21.0</td>
<td>VI</td>
<td>10.0</td>
<td>-11.0</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>25</td>
<td>CUMP</td>
<td>2.0</td>
<td>VI</td>
<td>4.0</td>
<td>2.0</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>26</td>
<td>CUMP</td>
<td>2.0</td>
<td>VI</td>
<td>2.0</td>
<td>0.0</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>27</td>
<td>VI</td>
<td>10.0</td>
<td>CUMP</td>
<td>6.0</td>
<td>-4.0</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>28</td>
<td>CUMP</td>
<td>7.0</td>
<td>CUMP</td>
<td>5.0</td>
<td>-2.0</td>
<td>0.73</td>
<td>0.80</td>
</tr>
<tr>
<td>29</td>
<td>CUMP</td>
<td>1.0</td>
<td>CUMP</td>
<td>10.0</td>
<td>9.0</td>
<td>0.85</td>
<td>0.90</td>
</tr>
<tr>
<td>30</td>
<td>VI</td>
<td>13.0</td>
<td>CUMP</td>
<td>11.0</td>
<td>-2.0</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>31</td>
<td>CUMP</td>
<td>17.0</td>
<td>CUMP</td>
<td>6.0</td>
<td>-11.0</td>
<td>0.74</td>
<td>0.87</td>
</tr>
<tr>
<td>32</td>
<td>CUMP</td>
<td>2.0</td>
<td>CUMP</td>
<td>6.0</td>
<td>4.0</td>
<td>0.54</td>
<td>0.59</td>
</tr>
<tr>
<td>33</td>
<td>VI</td>
<td>0.5</td>
<td>VI</td>
<td>0.5</td>
<td>0.0</td>
<td>0.68</td>
<td>0.73</td>
</tr>
<tr>
<td>Item</td>
<td>Pre-test method</td>
<td>Threshold (seconds)</td>
<td>Post-test method</td>
<td>Threshold (seconds)</td>
<td>Difference in thresholds (post-pre)</td>
<td>Pre-test</td>
<td>Post-test</td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>-------------------------------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>34</td>
<td>CUMP</td>
<td>17.0</td>
<td>CUMP</td>
<td>23.0</td>
<td>6.0</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>35</td>
<td>CUMP</td>
<td>2.0</td>
<td>CUMP</td>
<td>15.0</td>
<td>13.0</td>
<td>0.67</td>
<td>0.76</td>
</tr>
<tr>
<td>36</td>
<td>CUMP</td>
<td>6.0</td>
<td>CUMP</td>
<td>10.0</td>
<td>4.0</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>37</td>
<td>VI</td>
<td>10.5</td>
<td>CUMP</td>
<td>6.0</td>
<td>4.5</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td>38</td>
<td>VI</td>
<td>9.5</td>
<td>VI</td>
<td>8.0</td>
<td>1.5</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td>39</td>
<td>CUMP</td>
<td>8.0</td>
<td>CUMP</td>
<td>3.0</td>
<td>5.0</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>40</td>
<td>CUMP</td>
<td>4.0</td>
<td>VI</td>
<td>2.0</td>
<td>2.0</td>
<td>0.64</td>
<td>0.76</td>
</tr>
<tr>
<td>41</td>
<td>CUMP</td>
<td>12.0</td>
<td>CUMP</td>
<td>5.0</td>
<td>7.0</td>
<td>0.66</td>
<td>0.75</td>
</tr>
<tr>
<td>42</td>
<td>CUMP</td>
<td>13.0</td>
<td>CUMP</td>
<td>13.0</td>
<td>0.0</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>43</td>
<td>CUMP</td>
<td>2.0</td>
<td>CUMP</td>
<td>16.0</td>
<td>14.0</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td>44</td>
<td>CUMP</td>
<td>2.0</td>
<td>VI</td>
<td>4.5</td>
<td>2.5</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>45</td>
<td>CUMP</td>
<td>4.0</td>
<td>VI</td>
<td>4.0</td>
<td>0.0</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>46</td>
<td>CUMP</td>
<td>3.0</td>
<td>VI</td>
<td>4.0</td>
<td>1.0</td>
<td>0.51</td>
<td>0.60</td>
</tr>
<tr>
<td>47</td>
<td>CUMP</td>
<td>1.0</td>
<td>VI</td>
<td>19.0</td>
<td>18.0</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>48</td>
<td>VI</td>
<td>4.0</td>
<td>CUMP</td>
<td>2.0</td>
<td>2.0</td>
<td>0.80</td>
<td>0.87</td>
</tr>
<tr>
<td>49</td>
<td>VI</td>
<td>5.0</td>
<td>VI</td>
<td>3.5</td>
<td>1.5</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>50</td>
<td>CUMP</td>
<td>1.0</td>
<td>CUMP</td>
<td>5.0</td>
<td>4.0</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>51</td>
<td>CUMP</td>
<td>1.0</td>
<td>CUMP</td>
<td>21.0</td>
<td>20.0</td>
<td>0.66</td>
<td>0.78</td>
</tr>
<tr>
<td>52</td>
<td>VI</td>
<td>4.5</td>
<td>CUMP</td>
<td>3.0</td>
<td>1.5</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>53</td>
<td>CUMP</td>
<td>14.0</td>
<td>CUMP</td>
<td>7.0</td>
<td>7.0</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>54</td>
<td>VI</td>
<td>10.0</td>
<td>VI</td>
<td>7.0</td>
<td>3.0</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>55</td>
<td>CUMP</td>
<td>3.0</td>
<td>CUMP</td>
<td>6.0</td>
<td>3.0</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>56</td>
<td>CUMP</td>
<td>2.0</td>
<td>VI</td>
<td>0.5</td>
<td>1.5</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>57</td>
<td>VI</td>
<td>2.0</td>
<td>VI</td>
<td>2.0</td>
<td>0.0</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>58</td>
<td>VI</td>
<td>14.5</td>
<td>CUMP</td>
<td>15.0</td>
<td>0.5</td>
<td>0.66</td>
<td>0.80</td>
</tr>
<tr>
<td>59</td>
<td>VI</td>
<td>14.0</td>
<td>VI</td>
<td>8.0</td>
<td>6.0</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>60</td>
<td>VI</td>
<td>17.5</td>
<td>VI</td>
<td>14.0</td>
<td>3.5</td>
<td>0.21</td>
<td>0.29</td>
</tr>
<tr>
<td>61</td>
<td>VI</td>
<td>13.0</td>
<td>CUMP</td>
<td>26.0</td>
<td>13.0</td>
<td>0.27</td>
<td>0.40</td>
</tr>
<tr>
<td>62</td>
<td>CUMP</td>
<td>18.0</td>
<td>CUMP</td>
<td>14.0</td>
<td>4.0</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>63</td>
<td>CUMP</td>
<td>37.0</td>
<td>CUMP</td>
<td>24.0</td>
<td>-13.0</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>64</td>
<td>VI</td>
<td>14.5</td>
<td>CUMP</td>
<td>28.0</td>
<td>13.5</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>65</td>
<td>CUMP</td>
<td>15.0</td>
<td>CUMP</td>
<td>10.0</td>
<td>-5.0</td>
<td>0.58</td>
<td>0.69</td>
</tr>
<tr>
<td>66</td>
<td>VI</td>
<td>5.0</td>
<td>CUMP</td>
<td>3.0</td>
<td>-2.0</td>
<td>0.47</td>
<td>0.65</td>
</tr>
</tbody>
</table>

*Note. VI = visual inspection; CUMP = cumulative proportion

* Item difficulties were calculated as the percent of examinees who correctly responded to the item (based on observed data).

Item difficulty and item response time thresholds had a correlation of approximately -0.29 for both pre- and post-test.
Table 6

<table>
<thead>
<tr>
<th>Number of Cluster 3 courses completed</th>
<th>Frequency (%) of examinees</th>
<th>Cumulative % of examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>33 (8.5%)</td>
<td>8.5%</td>
</tr>
<tr>
<td>1</td>
<td>103 (26.5%)</td>
<td>35.1%</td>
</tr>
<tr>
<td>2</td>
<td>94 (24.2%)</td>
<td>59.3%</td>
</tr>
<tr>
<td>3</td>
<td>76 (19.6%)</td>
<td>78.9%</td>
</tr>
<tr>
<td>4</td>
<td>50 (12.9%)</td>
<td>91.8%</td>
</tr>
<tr>
<td>5</td>
<td>25 (6.4%)</td>
<td>98.2%</td>
</tr>
<tr>
<td>6</td>
<td>6 (1.5%)</td>
<td>99.7%</td>
</tr>
<tr>
<td>7</td>
<td>1 (0.3%)</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 7

<table>
<thead>
<tr>
<th>Number of missing values across all 136 variables within an examinee</th>
<th>Frequency (%) of examinees with this many missing values</th>
<th>% of examinees with this many missing values or fewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17 (4.4%)</td>
<td>4.4%</td>
</tr>
<tr>
<td>1</td>
<td>45 (11.6%)</td>
<td>16.0%</td>
</tr>
<tr>
<td>2</td>
<td>53 (13.7%)</td>
<td>29.6%</td>
</tr>
<tr>
<td>3</td>
<td>64 (16.5%)</td>
<td>46.1%</td>
</tr>
<tr>
<td>4</td>
<td>50 (12.9%)</td>
<td>59.0%</td>
</tr>
<tr>
<td>5</td>
<td>39 (10.1%)</td>
<td>69.1%</td>
</tr>
<tr>
<td>6</td>
<td>23 (5.9%)</td>
<td>75.0%</td>
</tr>
<tr>
<td>7</td>
<td>24 (6.2%)</td>
<td>81.2%</td>
</tr>
<tr>
<td>8</td>
<td>13 (3.4%)</td>
<td>84.5%</td>
</tr>
<tr>
<td>9</td>
<td>6 (1.5%)</td>
<td>86.1%</td>
</tr>
<tr>
<td>10</td>
<td>13 (3.4%)</td>
<td>89.4%</td>
</tr>
<tr>
<td>11</td>
<td>7 (1.8%)</td>
<td>91.2%</td>
</tr>
<tr>
<td>12</td>
<td>1 (0.3%)</td>
<td>91.5%</td>
</tr>
<tr>
<td>13</td>
<td>3 (0.8%)</td>
<td>92.3%</td>
</tr>
<tr>
<td>14</td>
<td>4 (1.0%)</td>
<td>93.3%</td>
</tr>
<tr>
<td>15</td>
<td>1 (0.3%)</td>
<td>93.6%</td>
</tr>
<tr>
<td>16</td>
<td>3 (0.8%)</td>
<td>94.3%</td>
</tr>
<tr>
<td>17</td>
<td>1 (0.3%)</td>
<td>94.6%</td>
</tr>
<tr>
<td>18</td>
<td>4 (1.0%)</td>
<td>95.6%</td>
</tr>
<tr>
<td>20</td>
<td>1 (0.3%)</td>
<td>95.9%</td>
</tr>
<tr>
<td>21</td>
<td>3 (0.8%)</td>
<td>96.6%</td>
</tr>
<tr>
<td>22</td>
<td>1 (0.3%)</td>
<td>96.9%</td>
</tr>
<tr>
<td>23</td>
<td>3 (0.8%)</td>
<td>97.7%</td>
</tr>
<tr>
<td>25</td>
<td>2 (0.5%)</td>
<td>98.2%</td>
</tr>
<tr>
<td>33</td>
<td>1 (0.3%)</td>
<td>98.5%</td>
</tr>
<tr>
<td>36</td>
<td>2 (0.5%)</td>
<td>99.0%</td>
</tr>
<tr>
<td>39</td>
<td>1 (0.3%)</td>
<td>99.2%</td>
</tr>
<tr>
<td>40</td>
<td>1 (0.3%)</td>
<td>99.5%</td>
</tr>
<tr>
<td>54</td>
<td>1 (0.3%)</td>
<td>99.7%</td>
</tr>
<tr>
<td>71</td>
<td>1 (0.3%)</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Table 8

Results of Independent-Groups T-Tests Between Low- and Adequately-Motivated Examinees on Auxiliary Variables

<table>
<thead>
<tr>
<th></th>
<th>Low-Motivated Group</th>
<th>Adequately-Motivated Group</th>
<th>M difference</th>
<th>95% CI of M difference</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test: SB item % correct</td>
<td>32 .61 .11</td>
<td>356 .70 .09</td>
<td>.10</td>
<td>0.06 – 0.13</td>
<td>&lt;.001</td>
<td>0.96</td>
</tr>
<tr>
<td>Post-test: SB item % correct</td>
<td>34 .67 .09</td>
<td>354 .76 .10</td>
<td>.09</td>
<td>0.06 – 0.13</td>
<td>&lt;.001</td>
<td>0.98</td>
</tr>
<tr>
<td>SAT math</td>
<td>47 564.26 70.95</td>
<td>326 587.42 64.58</td>
<td>23.17</td>
<td>3.10 – 43.23</td>
<td>.024</td>
<td>0.34</td>
</tr>
<tr>
<td>SAT verbal</td>
<td>45 557.56 62.49</td>
<td>326 576.01 69.18</td>
<td>18.46</td>
<td>-2.94 – 39.85</td>
<td>.091</td>
<td>0.28</td>
</tr>
<tr>
<td>Cluster 3 GPA</td>
<td>51 2.76 0.76</td>
<td>304 2.97 0.76</td>
<td>0.21</td>
<td>-0.02 – 0.43</td>
<td>.075</td>
<td>0.27</td>
</tr>
<tr>
<td># Cluster 3 courses</td>
<td>53 2.21 1.35</td>
<td>335 2.30 1.48</td>
<td>0.09</td>
<td>-0.33 – 0.52</td>
<td>.674</td>
<td>0.06</td>
</tr>
</tbody>
</table>

a The assumption of homogeneity of variances was assessed and satisfied for each of the variables tested.
Table 9

Descriptive Statistics for NW-9 Scores: Observed Data (Unaltered)

<table>
<thead>
<tr>
<th>Number of Cluster Three courses</th>
<th>Pre-test $M$ (SD)</th>
<th>Post-test $M$ (SD)</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45.42 (5.67)</td>
<td>47.97 (8.16)</td>
<td>46.70 (7.09)</td>
</tr>
<tr>
<td>1</td>
<td>44.29 (7.09)</td>
<td>48.03 (7.25)</td>
<td>46.16 (7.40)</td>
</tr>
<tr>
<td>2</td>
<td>45.37 (6.52)</td>
<td>48.32 (7.10)</td>
<td>46.85 (6.96)</td>
</tr>
<tr>
<td>3+</td>
<td>45.11 (6.63)</td>
<td>49.73 (6.93)</td>
<td>47.42 (7.16)</td>
</tr>
<tr>
<td>Marginal</td>
<td>44.98 (6.65)</td>
<td>48.79 (7.18)</td>
<td></td>
</tr>
</tbody>
</table>

*Note. $N = 388$*
Table 10

Descriptive Statistics for NW-9 Scores: Motivation Filtered Data

<table>
<thead>
<tr>
<th>Number of Cluster</th>
<th>Three courses</th>
<th>Pre-test $M (SD)$</th>
<th>Post-test $M (SD)$</th>
<th>Marginal $M (SD)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45.87 (5.21)</td>
<td>49.35 (6.17)</td>
<td>47.61 (5.93)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>45.36 (6.52)</td>
<td>49.40 (6.36)</td>
<td>47.38 (6.73)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>45.87 (6.14)</td>
<td>49.40 (6.24)</td>
<td>47.63 (6.42)</td>
<td></td>
</tr>
<tr>
<td>3+</td>
<td>45.90 (5.97)</td>
<td>51.10 (6.07)</td>
<td>48.50 (6.55)</td>
<td></td>
</tr>
<tr>
<td>Marginal</td>
<td>45.75 (6.07)</td>
<td>50.10 (6.22)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. n = 335*
### Table 11

**Descriptive Statistics for NW-9 Scores: MI Data (m=100)**

<table>
<thead>
<tr>
<th>Number of Cluster Three courses</th>
<th>Pre-test $M$ ($SD$)</th>
<th>Post-test $M$ ($SD$)</th>
<th>Marginal $M$ ($SD$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45.52 (5.78)</td>
<td>48.61 (7.14)</td>
<td>47.06 (6.63)</td>
</tr>
<tr>
<td>1</td>
<td>44.60 (6.99)</td>
<td>48.51 (6.84)</td>
<td>46.56 (7.17)</td>
</tr>
<tr>
<td>2</td>
<td>45.87 (6.43)</td>
<td>49.07 (6.24)</td>
<td>47.47 (6.52)</td>
</tr>
<tr>
<td>3+</td>
<td>45.55 (6.31)</td>
<td>50.30 (6.49)</td>
<td>47.92 (6.82)</td>
</tr>
<tr>
<td>Marginal</td>
<td>45.37 (6.48)</td>
<td>49.38 (6.60)</td>
<td></td>
</tr>
</tbody>
</table>

*Note. N = 388*
Table 12

*Descriptive Statistics for NW-9 Difference Scores by Post-Hoc Method*

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>M</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data (unaltered)</td>
<td>388</td>
<td>3.81</td>
<td>0.30</td>
</tr>
<tr>
<td>Motivation filtered data</td>
<td>335</td>
<td>4.35</td>
<td>0.27</td>
</tr>
<tr>
<td>MI data (m=100)</td>
<td>388</td>
<td>4.01</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Table 13

*Descriptive Statistics for NW-9 Difference Scores by Number of Cluster Three Courses and Post-Hoc Method*

<table>
<thead>
<tr>
<th></th>
<th>0 courses</th>
<th></th>
<th>1 course</th>
<th></th>
<th>2 courses</th>
<th></th>
<th>3+ courses</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>SE</td>
<td>n</td>
<td>M</td>
<td>SE</td>
<td>n</td>
<td>M</td>
</tr>
<tr>
<td>Observed data (unaltered)</td>
<td>33</td>
<td>2.55</td>
<td>1.10</td>
<td>103</td>
<td>3.74</td>
<td>0.61</td>
<td>94</td>
<td>2.95</td>
</tr>
<tr>
<td>Motivation filtered data</td>
<td>31</td>
<td>3.48</td>
<td>0.90</td>
<td>84</td>
<td>4.05</td>
<td>0.57</td>
<td>82</td>
<td>3.54</td>
</tr>
<tr>
<td>MI data (m=100)</td>
<td>33</td>
<td>3.09</td>
<td>0.94</td>
<td>103</td>
<td>3.91</td>
<td>0.56</td>
<td>94</td>
<td>3.19</td>
</tr>
</tbody>
</table>
Table 14

Mixed ANOVA Unstandardized Parameter Estimates and Standard Errors

<table>
<thead>
<tr>
<th>Effect</th>
<th>intercept</th>
<th>time(^a)</th>
<th>1 course(^b)</th>
<th>2 courses(^c)</th>
<th>3+ courses(^d)</th>
<th>1 course*time</th>
<th>2 courses*time</th>
<th>3+ courses*time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>SE</td>
</tr>
<tr>
<td>Observed data (unaltered)</td>
<td>46.78</td>
<td>0.37</td>
<td>1.73</td>
<td>0.17</td>
<td>-0.62</td>
<td>0.58</td>
<td>0.06</td>
<td>0.64</td>
</tr>
<tr>
<td>Motivation filtered data</td>
<td>47.78</td>
<td>0.35</td>
<td>2.03</td>
<td>0.16</td>
<td>-0.40</td>
<td>0.56</td>
<td>-0.15</td>
<td>0.56</td>
</tr>
<tr>
<td>MI data (m=100)</td>
<td>47.25</td>
<td>0.36</td>
<td>1.87</td>
<td>0.16</td>
<td>-0.70</td>
<td>0.55</td>
<td>0.22</td>
<td>0.57</td>
</tr>
</tbody>
</table>

\(^a\)Time was coded -1 for pre-test, 1 for post-test

\(^b\)Effect coded (1 course = 1; 2 or 3+ courses = 0; 0 courses = -1)

\(^c\)Effect coded (2 courses = 1; 1 or 3+ courses = 0; 0 courses = -1)

\(^d\)Effect coded (3+ courses = 1; 1 or 2 courses = 0; 0 courses = -1)
Table 15

*Tests of Mixed ANOVA Main Effects and Interaction*

<table>
<thead>
<tr>
<th>Effect</th>
<th>df&lt;sub&gt;effect&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>F</th>
<th>p</th>
<th>df&lt;sub&gt;effect&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>F</th>
<th>p</th>
<th>df&lt;sub&gt;effect&lt;/sub&gt;</th>
<th>df&lt;sub&gt;error&lt;/sub&gt;</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data (unaltered)</td>
<td>1</td>
<td>384</td>
<td>99.74</td>
<td>&lt;0.0001</td>
<td>3</td>
<td>384</td>
<td>0.85</td>
<td>0.47</td>
<td>3</td>
<td>384</td>
<td>2.26</td>
<td>0.08</td>
</tr>
<tr>
<td>Motivation filtered data</td>
<td>1</td>
<td>331</td>
<td>171.14</td>
<td>&lt;0.0001</td>
<td>3</td>
<td>331</td>
<td>0.85</td>
<td>0.47</td>
<td>3</td>
<td>331</td>
<td>2.54</td>
<td>0.06</td>
</tr>
<tr>
<td>MI data (m=100)</td>
<td>1</td>
<td>289217</td>
<td>140.99</td>
<td>&lt;0.0001</td>
<td>3</td>
<td>1.72E7</td>
<td>1.12</td>
<td>0.34</td>
<td>3</td>
<td>713690</td>
<td>2.12</td>
<td>0.10</td>
</tr>
</tbody>
</table>

*Note.* Degrees of freedom for the MI method pertain to the D<sub>1</sub> multiparameter test.
Table 16

**Pooled Means and Standard Errors For NW-9 Total Scores by Number of Imputations**

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th></th>
<th>Post-test</th>
<th></th>
<th>Difference (post–pre)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SE$</td>
<td>$M$</td>
<td>$SE$</td>
<td></td>
</tr>
<tr>
<td>MI data ($m=3$)</td>
<td>45.3909</td>
<td>0.3327</td>
<td>49.3505</td>
<td>0.3379</td>
<td>3.9596</td>
</tr>
<tr>
<td>MI data ($m=5$)</td>
<td>45.3768</td>
<td>0.3315</td>
<td>49.3660</td>
<td>0.3365</td>
<td>3.9892</td>
</tr>
<tr>
<td>MI data ($m=10$)</td>
<td>45.3747</td>
<td>0.3312</td>
<td>49.3680</td>
<td>0.3365</td>
<td>3.9933</td>
</tr>
<tr>
<td>MI data ($m=25$)</td>
<td>45.3785</td>
<td>0.3304</td>
<td>49.3771</td>
<td>0.3358</td>
<td>3.9987</td>
</tr>
<tr>
<td>MI data ($m=50$)</td>
<td>45.3726</td>
<td>0.3301</td>
<td>49.3796</td>
<td>0.3360</td>
<td>4.0070</td>
</tr>
<tr>
<td>MI data ($m=100$)</td>
<td>45.3750</td>
<td>0.3299</td>
<td>49.3808</td>
<td>0.3367</td>
<td>4.0059</td>
</tr>
</tbody>
</table>
Table 17

**Pooled Mixed ANOVA Parameter Estimates and Standard Errors by Number of Imputations**

<table>
<thead>
<tr>
<th>Effect</th>
<th>intercept</th>
<th>time</th>
<th>1 course</th>
<th>2 courses</th>
<th>3+ courses</th>
<th>1 course*time</th>
<th>2 courses*time</th>
<th>3+ courses*time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>B</td>
<td>SE</td>
<td>B</td>
</tr>
<tr>
<td>MI data (m=3)</td>
<td>47.239</td>
<td>0.357</td>
<td>1.844</td>
<td>0.165</td>
<td>-0.698</td>
<td>0.549</td>
<td>0.215</td>
<td>0.568</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI data (m=5)</td>
<td>47.244</td>
<td>0.357</td>
<td>1.854</td>
<td>0.161</td>
<td>-0.702</td>
<td>0.549</td>
<td>0.214</td>
<td>0.566</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI data (m=10)</td>
<td>47.246</td>
<td>0.358</td>
<td>1.859</td>
<td>0.159</td>
<td>-0.691</td>
<td>0.550</td>
<td>0.225</td>
<td>0.566</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI data (m=25)</td>
<td>47.253</td>
<td>0.358</td>
<td>1.865</td>
<td>0.157</td>
<td>-0.683</td>
<td>0.550</td>
<td>0.218</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI data (m=50)</td>
<td>47.253</td>
<td>0.358</td>
<td>1.868</td>
<td>0.157</td>
<td>-0.697</td>
<td>0.550</td>
<td>0.213</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI data (m=100)</td>
<td>47.254</td>
<td>0.358</td>
<td>1.867</td>
<td>0.157</td>
<td>-0.698</td>
<td>0.550</td>
<td>0.217</td>
<td>0.566</td>
</tr>
</tbody>
</table>
Table 18

*Tests of Mixed ANOVA Main Effects and Interaction by Number of Imputations*

<table>
<thead>
<tr>
<th>Effect</th>
<th>time</th>
<th>number of courses</th>
<th>time*number of courses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df$_{effect}$</td>
<td>df$_{error}$</td>
<td>F</td>
</tr>
<tr>
<td>MI data ($m=3$)</td>
<td>1</td>
<td>170.4</td>
<td>124.39</td>
</tr>
<tr>
<td>MI data ($m=5$)</td>
<td>1</td>
<td>1077</td>
<td>133.06</td>
</tr>
<tr>
<td>MI data ($m=10$)</td>
<td>1</td>
<td>3298.7</td>
<td>137.51</td>
</tr>
<tr>
<td>MI data ($m=25$)</td>
<td>1</td>
<td>45245</td>
<td>141.56</td>
</tr>
<tr>
<td>MI data ($m=50$)</td>
<td>1</td>
<td>122830</td>
<td>141.92</td>
</tr>
<tr>
<td>MI data ($m=100$)</td>
<td>1</td>
<td>289217</td>
<td>140.99</td>
</tr>
</tbody>
</table>

*Note.* Degrees of freedom for the MI methods pertain to the $D_1$ multiparameter test.
Table 19

*Between- and Within-Imputation Variance Estimates by Number of Imputations*

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Difference (post–pre)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_{between}$</td>
<td>$\sigma^2_{within}$</td>
<td>$\sigma^2_{between}$</td>
</tr>
<tr>
<td>MI data ($m=3$)</td>
<td>0.0023</td>
<td>0.1076</td>
<td>0.0015</td>
</tr>
<tr>
<td>MI data ($m=5$)</td>
<td>0.0015</td>
<td>0.1081</td>
<td>0.0012</td>
</tr>
<tr>
<td>MI data ($m=10$)</td>
<td>0.0013</td>
<td>0.1083</td>
<td>0.0011</td>
</tr>
<tr>
<td>MI data ($m=25$)</td>
<td>0.0010</td>
<td>0.1081</td>
<td>0.0010</td>
</tr>
<tr>
<td>MI data ($m=50$)</td>
<td>0.0008</td>
<td>0.1082</td>
<td>0.0009</td>
</tr>
<tr>
<td>MI data ($m=100$)</td>
<td>0.0007</td>
<td>0.1081</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Table 20 (continued on next page)

*Between- and Within-Imputation Variance Estimates by Number of Imputations*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Intercept</th>
<th>Time</th>
<th>1 course</th>
<th>2 courses</th>
<th>3+ courses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_{between}$</td>
<td>$\sigma^2_{within}$</td>
<td>$\sigma^2_{between}$</td>
<td>$\sigma^2_{within}$</td>
<td>$\sigma^2_{between}$</td>
</tr>
<tr>
<td>MI data ($m=3$)</td>
<td>0.0002</td>
<td>0.1270</td>
<td>0.0022</td>
<td>0.0244</td>
<td>0.0008</td>
</tr>
<tr>
<td>MI data ($m=5$)</td>
<td>0.0002</td>
<td>0.1272</td>
<td>0.0013</td>
<td>0.0243</td>
<td>0.0005</td>
</tr>
<tr>
<td>MI data ($m=10$)</td>
<td>0.0003</td>
<td>0.1275</td>
<td>0.0007</td>
<td>0.0244</td>
<td>0.0008</td>
</tr>
<tr>
<td>MI data ($m=25$)</td>
<td>0.0004</td>
<td>0.1275</td>
<td>0.0005</td>
<td>0.0241</td>
<td>0.0009</td>
</tr>
<tr>
<td>MI data ($m=50$)</td>
<td>0.0005</td>
<td>0.1276</td>
<td>0.0004</td>
<td>0.0241</td>
<td>0.0012</td>
</tr>
<tr>
<td>MI data ($m=100$)</td>
<td>0.0005</td>
<td>0.1277</td>
<td>0.0004</td>
<td>0.0243</td>
<td>0.0012</td>
</tr>
</tbody>
</table>
Table 20 (continued)

*Between- and Within-Imputation Variance Estimates by Number of Imputations*

<table>
<thead>
<tr>
<th>Effect</th>
<th>1 course*time</th>
<th>2 courses*time</th>
<th>3+ courses*time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_{\text{between}}$</td>
<td>$\sigma^2_{\text{within}}$</td>
<td>$\sigma^2_{\text{between}}$</td>
</tr>
<tr>
<td>MI data ($m=3$)</td>
<td>0.0004</td>
<td>0.0576</td>
<td>0.0004</td>
</tr>
<tr>
<td>MI data ($m=5$)</td>
<td>0.0004</td>
<td>0.0574</td>
<td>0.0002</td>
</tr>
<tr>
<td>MI data ($m=10$)</td>
<td>0.0009</td>
<td>0.0576</td>
<td>0.0005</td>
</tr>
<tr>
<td>MI data ($m=25$)</td>
<td>0.0014</td>
<td>0.0569</td>
<td>0.0011</td>
</tr>
<tr>
<td>MI data ($m=50$)</td>
<td>0.0013</td>
<td>0.0570</td>
<td>0.0016</td>
</tr>
<tr>
<td>MI data ($m=100$)</td>
<td>0.0013</td>
<td>0.0574</td>
<td>0.0013</td>
</tr>
</tbody>
</table>
Figure 1. The distribution of response times for item 5 of the NW-9 post-test administration.
Figure 2. The cumulative proportion correct function (solid line) and chance correct function (dashed line) for item 8 of the NW-9 pre-test administration.
Figure 3. The three mechanisms of missing data (left to right: MNAR, MAR, MCAR). Solid lines between observed variables indicate a non-zero relationship is present; dashed lines indicate a non-zero relationship may or may not be present; the absence of a line (dashed or solid) indicates a non-relationship, controlling for the other variables present.
Figure 4. Median item response times for each item of the NW-9, at both pre-test (light blue bars) and post-test (dark blue bars) administrations.
Figure 5. Item response time thresholds for each item of the NW-9, at both pre-test (light green bars) and post-test (dark green bars) administrations.
Figure 6. Overlapping distributions of examinees’ RTE scores at pre-test (red area) and post-test (blue area) administrations.
Figure 7. Average NW-9 scores by number of Cluster 3 courses completed at post-test, time, and post-hoc method.

Note. The vertical (Y) axis does not include the full range of NW-9 scores (i.e., 0-66); this was done in effort to better display the distinct lines.
Figure 8. Average NW-9 scores by number of Cluster 3 courses completed at post-test, time, and post-hoc method. 

Note. The vertical (Y) axis here does include the full range of NW-9 scores (unlike in Figure 7).