Fall 12-2019

Evaluation of Modern Missing Data Handling Methods for Coefficient Alpha

Katerina Matysova
University of Nebraska - Lincoln, kat.mat@huskers.unl.edu

EVALUATION OF MODERN MISSING DATA HANDLING METHODS FOR COEFFICIENT ALPHA

by

Katerina Matysova

A THESIS

Presented to the Faculty of

the Graduate College at the University of Nebraska

in Partial Fulfillments of Requirements

for the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor Rafael De Ayala

Lincoln, Nebraska

December, 2019

EVALUATION OF MODERN MISSING DATA HANDLING METHODS FOR COEFFICIENT ALPHA

Katerina Matysova, M.A.

University of Nebraska, 2019

Advisor: Rafael De Ayala

When assessing a certain characteristic or trait using a multiple item measure, quality of that measure can be assessed by examining the reliability. To avoid multiple time points, reliability can be represented by internal consistency, which is most commonly calculated using Cronbach's coefficient alpha. Almost every time human participants are involved in research, there is missing data involved. Missing data means that even though complete data were expected to be collected, some data are missing. Missing data can follow different patterns as well as be the result of different mechanisms. One traditional way to deal with missing data is listwise deletion, in which every observation with at least one missing value is discarded. Modern missing data techniques include multiple imputation and maximum likelihood estimation, which use the observed data to create an estimate for the missing values in order to utilize the whole sample size. The present study sought to examine the effect of missing data on coefficient alpha under certain conditions as well as to compare multiple imputation to listwise deletion in its effectiveness to handle missing data across those conditions. The results indicated that coefficient alpha is sensitive to numerous factors in the presence of missing data such as reliability level, sample size, missing data percentage, and missing data mechanism. As expected, there was little difference between listwise deletion and

multiple imputation when data were missing completely at random, but multiple imputation performed better when data were missing at random and missing not at random. While listwise deletion always underestimated the true reliability, multiple imputation only underestimated the true reliability when data were missing not at random.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

When assessing a certain characteristic or trait using a multiple item measure, researchers are often concerned with the quality of the measure. The quality can be determined by examining the validity and reliability of that measure. Obtaining the same result over multiple assessments of the same measure is defined as its reliability. It is the precursor to validity (i.e. validity of a measure cannot be determined without that measure being reliable) and the focus of this study.

To consider reliability from classical test theory perspective, it is important to look at the classical test theory model (i.e. $X = T + e$). In this model, the observed score X is composed of the true score T and an error term e. Since each of those terms is associated with a specific variance, reliability can be defined as the ratio of true score variance to observed score variance. The more observations vary from one assessment to the next, the lower will be the reliability of the measure and vice versa. Since the true scores can never be observed, reliability cannot be calculated, only estimated. Methods to estimate reliability will be discussed in more detail below.

Since reliability is a proportion, it can range between 0 and 1, while higher numbers indicate higher reliability. While it is evident that a higher reliability is desirable, it is important to consider the context in which it is being estimated. The attenuation paradox (Loevinger, 1954) describes the trade-off between reliability and validity. Maximization of reliability is only reasonable as long as it is not at the expense of validity. Standardized tests are usually recommended to have reliabilities of 0.7, 0.8, and 0.9 for low-stakes, medium-stakes, and high-stakes conditions respectively (Bonett, 2002).

The way reliability is estimated depends on the attribute of interest. Stability refers to the consistency of results over time. For example, a test can be administered at two different time points, with a certain time interval between them. In that case, reliability is the correlation between the test results of the two time points. When equivalence is of interest instead, two alternate versions of a measure can be used for assessment and reliability can be estimated by correlating their results. Both of those attributes require multiple assessments of a measure, which can be expensive and time consuming. A third attribute of reliability is internal consistency and only requires one measure to estimate reliability. It is based on the assumption that items testing the same construct should be correlated (Kimberlin & Winterstein, 2008). Ideally, when reliability is estimated, all three attributes are taken into account. Fortunately, researchers argue that when internal consistency indexes are calculated correctly, they accurately reflect stability and equivalence in addition to internal consistency (Wainer & Thissen, 1996).

When estimating internal consistency, the most evident conventional approach was to split the measure into two halves and calculate the correlation between them, referred to as the split-half method (Spearman, 1910; Brown, 1910). This method, however, reduces the sample size of the measure and underestimates reliability. Over time, researchers have improved upon the original idea, resulting in two internal consistency measures that are used today. Cronbach's coefficient alpha (Cronbach, 1951) is used with strictly unidimensional measures, while McDonald‘s coefficient omega (McDonald, 1999) relaxes that requirement. Despite both being considered equivalent

alternatives, coefficient alpha is more commonly reported in social science studies and will be used to estimate reliability in this study as well.

Cronbach's coefficient alpha is a function of all possible split-half coefficients for a measure, as well as a special case of the split-half reliability estimate (Cronbach, 1951). It is calculated based on the essentially tau-equivalent model, which assumes that that all items measure the same latent variable and that that they are measured on the same scale, but they can be measured with a varying degree of precision and have different error variances (Graham, 2006). In this case, precision refers to the equality of the strength of items. For example, a strongly worded item can result in a different response than a more weakly worded item. Coefficient alpha is not a test of unidimensionality, as researchers sometimes falsely believe, but rather a lower bound estimate of reliability (Miller, 1995).

Reliability, and hence also internal consistency, is not an intrinsic property of a measure, instead it depends on the sample variability in a specific case. This variability is influenced by multiple factors such as the number of items on a test, covariances between items, as well as a large sample are generally associated with higher reliability. There have been conflicting opinions on whether the number of response categories influences reliability, with the most popular being that an increase of reliability with the number of categories levels off after a certain point. Furthermore, there are factors that will not be discussed in this study, such as time restrictions or item selection, but which may also affect the reliability of a measure.

When estimating reliability, much like with other statistics, nonresponse is almost always involved. Nonresponse means that even though complete data were expected to

be collected, some data are missing. The data may be missing for an item, a measure, or in longitudinal studies, for a time point and follow different patterns and mechanisms. The three missing data mechanisms are used to determine whether values are missing at random or if there is a reason for their missingness, either observed or unobserved. They can have a large effect on sample statistics, especially because they are hard or even impossible to be identified.

The most common approach to handle the missing data is listwise deletion. However, deleting the whole observation if they have one or more missing values not only decreases the sample size, but it only returns unbiased results when data are missing completely at random. To deal with this problem, researchers have developed single imputation methods, which replace the missing values with predicted values. These methods range from simply using the mean to creating a regression equation and using the dependent variable as the predicted value.

Modern missing data handling methods, such as multiple imputation and maximum likelihood estimation also work when data is only missing at random. While maximum likelihood deals with the missing data during the model fitting procedure, multiple imputation creates complete data sets with imputed values before running the analysis. Even though an effort has been made to find methods to handle missing data when the data are missing not at random, those methods are still far from being convenient. Since both modern missing data handling methods tend to provide similar results (Schafer & Graham, 2002), this study focuses on multiple imputation only.

This study‘s objective is to examine the effect of missing data on coefficient alpha under certain conditions, such as varying sample size, missing data percentage, and missing data mechanism. In addition, multiple imputation is compared to listwise deletion in its effectiveness to handle missing data across the above conditions.

LITERATURE REVIEW

Quality of Measures.

Measurement uses the assignment of numbers to observations in order to quantify certain phenomena. In social sciences and other related fields, these phenomena such as intelligence or depression are often theoretical constructs that are not directly measurable. Therefore, psychometrics is concerned with the development of measures to quantify these constructs. A commonly used example of measuring attitudes, character, and personality traits is the Likert scale (Likert, 1932). Likert scales consist of multiple items measuring the same construct, in which each item has multiple, usually five or seven, response options. These options range from one extreme, such as “strongly agree”, to another extreme, such as “strongly disagree”. However, factors such as whether the positive or the negative option is on the left, choice of the measure, and reverse coding can influence the way participants respond to questions (Hartley, 2013).

The goal of measurement is to capture the true underlying characteristics to represent the construct of interest as precisely as possible. Therefore, the quality of measures should be assessed every time they are used for assessment. Key indicators of the quality of measures are reliability and validity (i.e., Kimberlin & Winterstein, 2008). Validity is an assessment of the extent to which a measure is accurately assessing the phenomenon that it is intended to represent (Cronbach & Meehl, 1955). Reliability is the consistency of the measure, the extent to which a measure would provide the same results in successive administrations of that measure, with different persons, on different occasions, under different conditions (Drost, 2011). In other words, it is the magnitude of the error of measurement (Cronbach, 1951).

When the measure consists of multiple items, reliability is also the extent to which all of the items measure the same phenomenon. It is impossible to determine whether a measure accurately measures what it intends to if it is not consistent from one administration to the next. Hence, a measure must be reliable before validity can be assessed.

Measurement Models.

In the ideal world, our observed scores, *X*, would be measured without measurement error. However, it is more realistic to believe that all observed scores contain error. The classical test theory model, $X = T + e$, reflects this latter case. This true score model assumes that the observed score *X* consists of true score *T*, which is the underlying score that would be obtained if there was no error in the measurement, plus an error term *e* (Downing, 2004). The true score is latent, meaning that it is unobserved. Since the true score is the average of all the observed scores a person would obtain if their score was assessed an infinite number of times, each observed score will contain a certain amount of error (i.e., $e = X - T$). This error can be positive or negative, but it does not bias the measure in a systematic way, so that the mean of the error term is expected to be 0 (Allen & Yen, 1979).

In the classical test theory model, reliability is defined as the ratio of true score variance to total observed score variance (Cronbach, 1951). There are no parameters given to divide the observed variance into true score and error variance. Therefore, the division can be done an infinite number of ways (Graham, 2006). It is impossible to estimate reliability, unless the measure is made of multiple items. Even with multiple

items, there are still numerous ways to partition the variance. In order to estimate reliability of items in a measure, it is necessary to make assumptions about the relationship of those items. This results in a number of measurement models with different requirements for the data (Graham, 2006).

The *parallel model* is the most restrictive model. It assumes that all of the items are exactly the same (i.e., they measure the same latent variable) are measured on the same scale, have the same degree of precision, and the same amount of error (Raykov, 1997a, 1997b). A precise measure is one in which values of the measured items are close together, while in an imprecise measure the values are widely spread out. For example, a 5-point Likert scale assessing anxiety with answer options from “strongly agree” to “strongly disagree” contains two items, one being “I feel nervous sometimes” and the second being “I almost always feel nervous”. Even though those two items measure the same latent variable and are on the same scale, they will possibly result in differences in precision, because one is more strongly worded than the other. With each item *j* for individual *i*, the parallel model can be applied to the classical test theory model, so that (Graham, 2006)

$$X_{ij} = T_i + e_i \tag{2.1}$$

The *tau-equivalent model* is less restrictive, as it allows for a different error variance for each item. As in the parallel model, the tau-equivalent model also assumes that the items measure the same latent variable, are measured on the same scale, and have the same degree of precision (Raykov, 1997a, 1997b). Therefore, all unique variance associated with an item is due to the error variance, so that (Graham, 2006)

$$X_{ij} = T_i + e_{ij} \quad (2.2)$$

The *essentially tau-equivalent model* is less restrictive than the tau-equivalent model, as it relaxes the assumption of equal precision. Essential tau-equivalence still assumes that that all items measure the same latent variable and that that they are measured on the same scale, but they can be measured with a varying degree of precision (Raykov, 1997a). With α being an additive constant for each individual item, this model can be mathematically represented as (Graham, 2006)

$$X_{ij} = (\alpha_j + T_i) + e_{ij} \quad (2.3)$$

The congeneric model is the least restrictive model, because it only requires the assumption that each item measures the same latent variable. The items could be measuring different scales, at different degrees of precision, and with different amounts of error (Raykov, 1997a). The congeneric model assumes a linear relationship between true scores and observed scores, with not only an additive constant, but also a multiplicative constant β, so that (Graham, 2006)

$$X_{ij} = [\alpha_j + \beta_j (T_i)] + E_{ij} \quad (2.4)$$

Reliability.

Measurement scales almost always include some amount of measurement error. As outlined in the classical test theory model, the true score can never be actually observed. Instead, it is the average score a person would obtain if they took a measure an infinite number of times (Allen & Yen, 1979). Consequently, a person's observed score will vary around the true score to some degree, so that all of the terms in the classical test theory model have their own specific variance. Based on the classical test theory model,

reliability is defined as the ratio of true score variance (σ^2_{True}) to total observed score variance (σ^2_{Total}) (Streiner, 2003):

$$Reliability = \frac{\sigma^2_{True}}{\sigma^2_{Total}} \quad (2.5)$$

Reliability provides a tool of estimating the amount of measurement error in assessments (Downing, 2004). As a proportion of variances its range is from 0 to 1, in which 0 is no reliability at all and 1 is perfect reliability. As discussed below in this study, an estimate of reliability may, however, be outside of that range.

The size of the optimal reliability for a given measure depends on what it is intended to measure. The dependency of the reliability coefficient on the sample makes a determination of objective guidelines difficult. Similarly to the low-stakes, medium-stakes, and high-stakes guidelines for cognitive measurements mentioned above, some recommendations have been suggested for psychological tests as well. Steinborn, Langner, Flehmig, and Huestegge (2017) proposed that a reliability estimate of 0.9 or higher is considered high, a reliability estimate between 0.8 – 0.89 is sufficient, and a reliability below 0.8 is problematic.

The implication is that higher reliability is better, but this is only the case to a certain point. Loevinger (1954) described the *attenuation paradox* as the trade-off between reliability and validity. A maximization of reliability inevitably results in a decrease of validity. For instance, if half of the observations on a test were perfectly 0 (i.e., half of the participants made zero scores) and the other half of the observations were perfectly 1 (i.e., half of the participants made perfect scores), then the reliability

coefficient would be 1. However, this is not desirable for obvious reasons. Therefore, maximization of reliability is only desirable as long as it is not at the expense of validity.

Reliability is not an intrinsic property of a measure. Rather, a measure's reliability varies across different circumstances. Reliability is just as dependent on the sample variability in a specific case as are other statistics of the measure, such as item difficulty or discrimination. Therefore, reliability estimators reflect a characteristic of test scores, not the test itself (Yin & Fan, 2000). Moreover, a test's scores are not determined to be either reliable or unreliable in all conditions, but determined to be reliable or unreliable for a particular sample. The reason for this is that reliability coefficients are parameter estimates instead of sample statistics. Therefore, they will always include sampling error to some degree and will change depending on the specific sample used to estimate reliability (Streiner, 2003).

Unfortunately, only observed scores are available when a measure is administered. It is not possible to determine how the observed score variance is divided into true score variance and error variance. Hence, reliability as the proportion of true score variance to total observed score variance cannot be directly measured, but has to be estimated instead. There are several approaches to estimating reliability.

Reliability Attributes.

Stability is one of three distinct attributes of reliability outlined by Drost (2011). It refers to the consistency of results across time. It can be measured by administering a test at two different points in time and correlating the results with each other. The interval between the measurements should be long enough to prevent the results from the first

administration to influence the results from the second administration (Kimberlin & Winterstein, 2008), but not so long that important resources such as time and money are wasted and results are potentially biased because of maturation between time points. When stability is the attribute of interest, reliability is estimated using the test-retest approach. In this case, the same measure is administered twice after a period of time and the correlation between the scores of the two assessments is calculated. One disadvantage of this approach is the need for multiple time points, resulting in a longitudinal design. Furthermore, depending on the length of the time period between the two assessments, the results at the first time point could influence the results at the second time point and bias the reliability estimate.

The second attribute *equivalence* refers to the test being used by different administrators or alternate measures being administered. When two or more judges evaluate a performance, then the level of agreement between their evaluations, referred to as the inter-rater reliability estimate, determines equivalence. Another method to evaluate equivalence is to administer two alternative forms of the same test and calculate the correlation between them.

Internal consistency is the third reliability attribute, which measures the equivalence of individual items on the same test. Researchers frequently use 'internal consistency' interchangeably with 'homogeneity' (Heale & Twycross, 2015). However, it is important to make a distinction between those terms to correctly estimate reliability (i.e., Cronbach, 1951). This distinction will be discussed below.

Internal consistency is based on the assumption that items measuring the same construct should be correlated (Kimberlin & Winterstein, 2008). Estimates of reliability measuring internal consistency have the advantage that only one admission of the measure is needed and are therefore used most frequently. Examples of internal consistency estimates are discussed below.

Even though the categorization of reliability into stability, equivalence, and internal consistency suggests different ways to estimate reliability, Wainer and Thissen (1996) argue that when internal consistency indexes are calculated correctly, they accurately reflect stability, such as test-retest correlation; as well as equivalence, such as correlation between alternate forms. For example, consider a measure A with a specific mean covariance between items. Then we sample from a pool of items with the same mean covariance as measure A without replacement to create two new tests (tests B and C) of the same length. The reliability estimate between measures B and C provides the coefficient of equivalence. When this process is repeated over and over, their mean would be the internal consistency of measure A (Cronbach, 1951).

Internal Consistency Estimates.

The definition of reliability is the extent to which a measure would provide the same results in successive administrations of that measure. Therefore, the most evident conventional approach to estimate internal consistency was to split the measure into two halves and calculate the correlation between them, referred to as the *split-half method* (Spearman, 1910; Brown, 1910). However, splitting the measure shortens it to half of its length, causing an underestimate of reliability (Streiner, 2003).

Spearman (1910) and Brown (1910) simultaneously developed a method, now referred to as *Spearman-Brown Prediction Formula,* which converts the split-half correlation into an estimate of reliability that takes into account that the correlation was calculated using only half of the test. A disadvantage of the Spearman-Brown Prediction Formula is the numerous ways to split each measure with each resulting in a different reliability coefficient. It also produces a biased estimate of reliability, which could either be an over- or an underestimate based on how the measure is split. This is due to the assumption that each added item increases reliability by the same amount (Kuder & Richardson, 1937).

Kuder and Richardson (1937) developed a formula to estimate reliability for binary items, referred to as *Kuder-Richardson Formula 20,* which is based on variance instead of correlation between items. Specifically, it provides a reliability estimate that is based on the number of items, the probability of not passing or passing an item (i.e., a coding of "0" and "1," respectively), and the overall variance of the measure. A second version is the *Kuder-Richardson Formula 21*, which is similar to the Kuder-Richardson Formula 20 but assumes that all items are equally difficult. However, the limitation of both Kuder-Richardson Formulas is that they can only be used with binary items. Cronbach (1951) solved that problem with his *coefficient alpha*, which can be used with continuous data, partial credit, and Likert scale (Zhang & Yuan, 2016).

Cronbach's coefficient alpha is calculated under the assumption that the measure is unidimensional and hence that the item covariances are the same. Therefore, it is a lower bound rather than an exact representation of reliability in cases where a measure

may be multidimensional. McDonald's (1999) *coefficient omega* relaxes this assumption and can be used with measures that are not strictly unidimensional.

While Cronbach's alpha and coefficient omega are both used as equivalent alternatives, Cronbach's alpha has been the most commonly used measure in numerous areas of research (Cortina, 1993) and is often reported with the use of measures that consist of multiple items (Downing, 2004; Zhang & Yuan, 2016).

Cronbach's Coefficient Alpha.

Cronbach's coefficient alpha is a function of all possible split-half coefficients for a measure, as well as a special case of the split-half reliability estimate (Cronbach, 1951). If a measure was split in half in all-possible ways, then the average between all of the split-half estimates is coefficient alpha. Reliability is defined, as previously mentioned, by the proportion of true score variance to observed score variance. Cronbach's coefficient alpha takes into account the number of items, so that the coefficient is the ratio of interitem covariance to total variance multiplied by *k*/(*k*-1) (Cronbach, 1951). It is estimated by the following formula (Streiner, 2003):

$$\frac{k}{k-1}\left(1-\frac{\sum_{j=1}^{k} var(x_j)}{var(x_0)}\right) \tag{2.6}$$

In which *k* refers to the number of items on the measure of interest, $var(x_j)$ refers to the variance associated with a specific item *j*, and refers to the total observed score variance. It can also be written as:

$$\alpha = \frac{k * cov(x_j, x_m)}{var(x_j) + (k-1) * cov(x_j, x_m)} \tag{2.7}$$

in which *j* and *m* are items with an average interitem covariance of $j \neq m$. To calculate the average interitem covariance, the correlation between each pair of item is calculated. For

example, if a measure had three items, then the correlation between item 1 and item 2, between item 2 and item 3, and between item 1 and item 3 is calculated. All of those correlations are averaged across to obtain the average interitem covariance. Therefore, Cronbach's alpha will increase when correlations between items increase. This means that if a measure would simply ask the same question over and over, assuming no measurement error, then coefficient alpha would equal 1 (Zhang & Yuan, 2016).

Technically, since reliability is the proportion of variance in the observed scores attributable to the total variance, its range is between 0 and 1. However, it is possible in some cases to obtain a negative estimate of reliability (e.g., a negative coefficient alpha). This could be the case if either the reverse questions were not reverse coded, leading to negative correlations between items (Streiner, 2003), or if there are different constructs that items belong to, leading to the variance of the individual items exceeding their shared variance (Henson, 2001). Both of those cases indicate that the items are not measuring what they are intended to, which points to problems with the original scale.

Coefficient alpha is calculated based on the essentially tau-equivalent model, which assumes that that all items measure the same latent variable and that that they are measured on the same scale, but they can be measured with a varying degree of precision and have different error variances (Graham, 2006). As a result, a common misconception between researchers is that coefficient alpha is a measure of unidimensionality (i.e., Cortina, 1993; Schmitt, 1996). To show that this assumption is wrong, it is important to define the difference between *internal consistency* and *homogeneity* (Cortina, 1993). While internal consistency refers to how a set of items is interrelated, homogeneity refers

to the unidimensionality of those items. Internal consistency is a necessary, but not sufficient condition for homogeneity. For instance, a set of items can be relatively interrelated and still represent multiple dimensions. On the contrary, homogeneity is necessary to accurately estimate reliability using coefficient alpha. Since coefficient alpha is a measure of item interrelation, it underestimates reliability when homogeneity is not present. Hence, coefficient alpha is not a test of unidimensionality, but rather a lower bound estimate of reliability (Miller, 1995).

Factors Influencing Internal Consistency.

There are numerous factors that can influence the internal consistency of a measure. Coefficient alpha is simply a function of the number of items, item variances, and covariances between items. These components can be used to determine the influences on internal consistency.

Test length, or the number of items in a measure, affects the internal consistency directly. There is a strong positive correlation between the number of items and internal consistency, as measured by multiple reliability coefficients (Javali, Gudaganavar, & Raj, 2011). That is, when other factors stay the same, internal consistency of a measure increases as the length of a scale increases. For example, a scale with more than 14 items will have a coefficient alpha of about .70, even if the correlations between items are only .30 (Cortina, 1993).

Sample size (i.e., the number of observations available for a measure) impacts the item variances of that measure in that a larger sample size generally results in smaller variance. Since the total variance is in the denominator of the mathematical

representation of reliability, a smaller total variance results in larger reliability. Sample size is not only one of the biggest challenges of a study's design because finding a balance between too small and too large is difficult (Bonett, 2002), but also has an effect on the reliability of the measures used. While Yurdugül (2008) found that for an unbiased estimate of reliability using coefficient alpha a sample size of 30 is sufficient with a large first eigenvalue and 100 is adequate with a smaller first eigenvalue, Ercan, Yazici, Sigirli, Ediz, and Kan (2007) argue that sample size is not important for coefficient alpha and estimates are stable even with very small sample sizes.

Number of response categories is a factor with different opinions about its influence on reliability. As Lissitz and Green (1975) point out, there have been numerous conflicting opinions on this topic, including the 7-point scale as the optimal number, independence of response category number and reliability, and a positive relationship between number of response categories and reliability. In their simulation, Lissitz and Green (1975) found that there is an increase in reliability with an increase in the number of response categories up to five categories. With more categories, the effect levels off. It is also important to note that the optimal number of response categories is dependent on the interests and objectives of a specific study more than on a recommended standard (Lissitz & Green, 1975).

Covariances between items represent the degree to which the items are related to each other, in other words it is the amount of shared variance between them. An item's communality is the extent to which each item correlates with at least one other item on the scale. As the correlations between items increase, coefficient alpha also increases

(Cortina, 1993). As mentioned above, asking the exact same question over and over in one scale would result in an coefficient alpha of 1 for that scale (Zhang & Yuan, 2016).

Symonds (1928) described 25 factors that influence reliability of measures in his paper. Most of those factors are concerned with the stability of the measure, rather than the internal consistency, and will not be discussed further in this paper. However, it is important to note that in addition to those mentioned above there are several different factors that can influence the reliability of a measure and to keep them in mind when talking about reliability. Some examples of those factors are time restrictions, item difficulty and discrimination, instructions, item selection, and scorer reliability (Symonds, 1928).

Missing Data.

Almost every time human participants are involved in research, there is missing data involved. With surveys and other measures that involve multiple items, an important type of missing data is nonresponse. Nonresponse means that even though complete data were expected to be collected, some data are missing. Nonresponse of one participant can apply to the complete measure or only certain parts of it.

Types of Nonresponse.

In survey methodology, data can be missing due to *unit nonresponse*, when no data are collected for the participant at all for that specific measure. For instance, this is due to the participant not being at home or refusing to participate in the study. In *item nonresponse* participants could answer parts of the survey, but leave some of the questions blank. This occurs because the participant does not know the answer, refuses to

answer, or a processing error such as wrong data entry, interviewer error, computer-assisted questionnaire programming error, or post-survey processing decisions which exclude parts of the data (Cuesta & Fonseca, 2014). In longitudinal studies, a participant may be present for some of the data collection time points, but may be missing for others, resulting in *wave nonresponse*. For example, attrition occurs when participants leave the study and do not return. This is the case when participants move away or with older populations, when participants die while the study is in progress.

Missing Data Patterns.

When item nonresponse is examined for the whole sample, certain patterns of nonresponse can be observed based on the location, or distribution, of the missing values. The pattern of missing data provides information about the configuration of observed and missing values in the data (Enders, 2010).

The *univariate pattern* shows a structure where responses on only one variable are missing. For instance, this is the case when an item that does not apply to a number of participants or when an exam question is very difficult.

The *unit nonresponse pattern* can be observed when some items are complete and other items have missing values for a subset of the participants. For instance, when a question asks whether participants smoke or not and then follows up with smoking behavior questions only for those who answered yes.

The *monotone pattern* is typically associated with a longitudinal design. As the study goes on, an increasing number of observations are missing for the measures administered later in time. This pattern resembles a staircase and is usually due to

participants who drop out of the study and never return. In terms of scale measures, this is the case for measures with time restrictions. The items at the end of the measure have an increasing amount of missingness.

The *general pattern* is most commonly observed and shows a seemingly random distribution of missing values throughout the data. The random pattern describes the location rather than the reason for missingness, so the data could still be systematically missing (i.e., with a certain missing data mechanism).

The *planned missing pattern* is used to reduce the number of items answered by each individual, while still collecting data on a large amount of items. This is the case when participants are divided into blocks and only respond to a certain portion of the items.

The *latent variable* pattern only applies to latent variable analyses such as structural equation models. Missing data follow this pattern when values for the indicator variables are available, but the latent variables are missing for the entire sample. Missing data algorithms can be used in that case, even though latent variable models may not be seen as missing data problems.

Missing Data Mechanisms.

Missing data mechanisms provide information about the relationship between the missing data and the variables in the model, if there is a relationship at all. These mechanisms were first described by Rubin (1976) and have been continuously used in research ever since. However, these mechanisms are only theoretical, since it is very difficult to impossible to find out which missing data mechanism is present in the data.

Therefore, researchers can rarely be sure about the true reason for missing data (Enders, 2010).

The three mechanisms are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). As described in Enders (2010), the complete data Y_{com} consists of two parts, the observed data Y_{obs} and the missing data Y_{mis}. The probability distribution p for each of the missing data mechanisms shows how the missing data indicator R depends on Y_{com}, Y_{obs}, Y_{mis}, and a parameter ϕ that describes the relationship between R and the data.

The three missing data mechanisms described by Rubin (1976) can be further described by whether they result in *ignorable* or *nonignorable* missingness. When estimates calculated from the data set are unbiased even in the presence of missing data, missingness is ignorable and no further adjustment is necessary. However, if the missing data are nonignorable, the estimates from the data set will be biased and need to be adjusted.

Missing data that follows the MNAR pattern are *nonignorable*, because the variable that causes the missingness is not measured. In this case, the absence of data on the outcome variable is related to the value on that variable. For example, in a study conducted on weight loss, participants who gain weight over the course of the study may be more likely to drop out of the study or not report their weight. Another example at the item level, poorly worded or confusing items may be intentionally omitted by participants and therefore considered MNAR (Parent, 2013). For example, items that ask about feelings towards a husband or boyfriend may be intentionally omitted by women in same-

gender relationships. If sexuality is not assessed in the survey, then there is no way to find out why data points are missing and the type of missingness is classified as MNAR. The distribution of MNAR can be written as:

$$p(R|Y_{obs}, Y_{mis}, \phi) \tag{2.8}$$

Data that are MAR show missingness due to a variable that is included in the dataset. In other words, the probability that data are missing depends on observed data and not on missing data. For example, when there is a speed limit on an exam people with slower reading speed are more likely to have some items missing. As long as reading speed is included as a variable, the missing data mechanism is considered MAR. As another example and using the previous item-level example, for the missing data to be MAR, sexuality needs to be assessed by the survey. When data are missing due to a MAR mechanism, modern missing data techniques, such as maximum likelihood estimation (MLE) and multiple imputation (MI) are most widely used. For MAR, the distribution can be described as:

$$p(R|Y_{obs}, \phi) \tag{2.9}$$

The MCAR mechanism is a special case of MAR and requires that the reason for why data are missing is completely unrelated to the study. As an example, this would be the case in the unit nonresponse and the wave nonresponse type of missingness if participants got sick on the day of data collection. In the item nonresponse case this might happen when a participant does not see a question or a machine malfunctions when

administering a survey on the computer. Similar to MAR, when data are MCAR the assumptions are met so that modern missing data techniques may be used to handle the missing data. MCAR is also the only mechanism that can be potentially detected, by using a multivariate extension of the t-test, as proposed by Little (1988). The distribution of MNAR is:

$$p(R|\phi) \tag{2.10}$$

Of course, preventing the occurrence of missing data is best. If there are no missing data to begin with, no methods are needed to adjust for them. Parent recommends ensuring that all items are applicable to all respondents. Furthermore, pilot studies can help to determine missing data patterns. Since MNAR is the data mechanism that is most difficult to deal with after the data are collected, it is the most important one to prevent. Therefore, it is very important to include questions that are related to the missing values to at least transform the MNAR missing data into MAR. For instance, as described in the previous item-level example, including a question about sexuality changes missing data on questions relating to a boyfriend for participants in same-sex relationships from MNAR to MAR.

Dealing with Missingness.

Most statistics, including coefficient alpha, would be impossible to calculate with an incomplete dataset. Therefore, either the dataset must be reduced, until sufficient information is reached, or the missing values must be imputed. The most common ways, but also the least effective, are *listwise deletion* and *pairwise deletion*. Listwise deletion

results in the deletion of each case that contains missing data, including the variables that do have observed values. This can result in an elimination of a large number of cases, wasting resources and causing lower statistical power (Enders, 2010). Pairwise deletion only deletes cases when they have missing data on one or more of the variables used for the statistic calculated, such that different statistics are based on different sample sizes. The problem with using these methods is that they assume the MCAR mechanism and will produce biased estimates if that assumption is violated (Enders, 2010). Even though these methods, most commonly listwise deletion, are still the default option in most statistical programs, they are being increasingly criticized in methodological literature and are slowly falling out of favor (Little & Rubin, 2002).

To prevent the problem of sample size reduction and the assumption of missing data being MCAR, another option is to impute the missing values based on the observed values. Imputing values replaces the missing values with a value that is calculated from the rest of the data that are complete. The decision to impute values has to make sense. For example, if a participant answers “no” to the question “Do you smoke?”, then it does not make sense to impute that participant’s answers on smoking behavior. Traditional methods of imputing missing values rely on single imputation, where each value is only imputed once. In more modern missing data handling methods values are imputed multiple times, resulting in variability of the imputed values.

Single Imputation Methods.

The most basic single imputation is *arithmetic mean imputation* (Wilks, 1932); that is, inserting the mean of the variable into all of the values that are missing. This

approach is very straightforward, but the problem is that imputing values from the center of the distribution decreases variability, thereby resulting in a severe distortion of the results even when the data are MCAR (Enders, 2010). For example, a decrease in standard error results in a larger test statistic as well as a higher probability of rejecting the null hypothesis. This imputation method is heavily criticized in methodological research (e.g., Enders, 2003; Schafer & Graham, 2002).

A second option is *regression imputation* (Buck, 1960), with the idea that a regression model based on variables being correlated with each other. The observed variables can be used to predict the missing variables, so that the regression model includes observed variables as predictor variables and the missing variable as the outcome variable. Since regression is used to make predictions, this method intuitively makes sense. However, the imputed values will fall exactly on the regression line, which causes an overestimation of the correlations and R^2, even when the data are MCAR (Enders, 2010).

Stochastic regression imputation was introduced to solve this overestimation. Stochastic regression imputation works the same way as regression imputation, but includes an error term into the predictions. This error term is randomly drawn from a normal distribution with a mean of zero and variance equal to the residual variance from the regression equation. Of all the single imputation methods, stochastic regression is the

only one that can provide unbiased parameter estimates under the MAR missing data mechanism (Enders, 2010).

Furthermore, the *hot-deck imputation* (Roth, 1994; Myers, 2011) is used to replace a missing value with a value that was obtained from other respondents. For example, the missing value can be replaced by a value randomly drawn from all of the other observed values, or from a subset of respondents who scored similarly as the respondent with the missing value on a set of matching variables (Enders, 2010).

Similarly, *cold-deck imputation* also uses a donor value to replace the missing value in the dataset (Shao 2000). However, in this case the value is obtained from anything other than reported values on the same item in the current data set (e.g. values from a previous item).

Averaging the available cases (Enders, 2010) is specific to multiple-item questionnaires measuring one construct. Researchers are often interested in a scale score (i.e., the average or the sum of the responses) rather than the individual responses. To obtain a complete scale score without throwing away cases, the average of available cases is used.

Last observation carried forward (Enders, 2010) can be used in longitudinal designs where participants have missing data for some of the time points, but not for others. In that case, the last observed value is used to replace the missing value.

Modern Missing Data Methods.

Compared to single imputation, modern missing data methods do not have to make the assumption that data are MCAR, but only MAR. They differ from listwise and

pairwise deletion in that they can use all of the observed data, without reducing the sample size. The two most successful methods that methodologists currently consider as state of the art are maximum likelihood estimation within the expectation-maximization algorithm and multiple imputation (Schafer & Graham, 2002). A big distinction between the two missing data techniques is that multiple imputation creates imputed values before running the analysis, while likelihood methods deal with the missing data during the model-fitting procedure. Overall, the two approaches have been shown to produce similar results, if the same data analysis model is used (Schafer & Graham, 2002). The increasing focus on missing data methods has been part of the reason why these methods are slowly becoming common procedure. Most statistical programs include maximum likelihood estimation and multiple imputation as missing data procedures in their packages.

Maximum Likelihood Estimation for Handling Missing Data.

Originally, MLE is a process that has not been specifically developed for the purpose of dealing with missing data. Instead, it is a way to estimate population parameters based on their likelihood from a certain sample (Enders, 2010). Even though using maximum likelihood estimation in the presence of missing data is not a new concept (e.g., Edgett, 1956), it has only recently been implemented into statistical programs and become accessible to researchers with limited methodological background (Schafer & Graham, 2002). In literature, maximum likelihood data handling is often also referred to as *full estimation maximum likelihood* (FIML) or *direct maximum likelihood*.

This method can now be found in most statistical packages, even though the

utilization may still be limited or only available for certain analyses (Enders, 2010).

The first step of maximum likelihood estimation (MLE) is to either determine the distribution of the variable or variables or assume a particular distribution. Typically, the distribution is assumed to be normal or multivariate normal in the multivariate case. From this distribution a probability density function is developed that indicates the probability of obtaining a certain value for that specific distribution, also called the likelihood. To identify parameters from a sample of data, the likelihood of individual cases is multiplied, resulting in a likelihood for the sample. For a more useful metric, the natural logarithm of likelihood values is used instead of individual values (Enders, 2010).

Assuming that the likelihood is a differentiable function, then the first and second derivatives are used to maximize the probability density function. The first derivative is set to 0 to determine the locations of the extrema (i.e., one or more minima and/or maxima). If the second derivative at that location is negative, the extrema is known to be a maximum of the function, while a positive second derivative indicates a minimum of the function. The goal is to identify population parameters with the highest likelihood (i.e. the global maximum of the log-likelihood function) of producing the particular sample (Enders, 2010). However, estimating more complex models can require a collection of equations with one or more unknowns. Since solving for unknown parameter values in a set of equations can be complex, an iterative optimization algorithm can be used with advanced applications of maximum likelihood estimation (Enders, 2010). The iterative optimization algorithm used with missing data is the expectation-maximization algorithm described below.

The advantage of using MLE to handle missing data is that it yields unbiased parameter estimates under the MAR missing data mechanism. However, it may still result in biased parameter estimates when the missing data are MNAR (Enders, 2010).

Expectation-Maximization Algorithm.

The expectation-maximization (EM) algorithm is an iterative optimization algorithm performed to estimate unknown parameters for data with missing values (Dempster, Laird, & Rubin, 1977). The expectation-maximization algorithm does not "fill in" the missing data, but rather maximizes the complete log-likelihood function to estimate the unknown parameter. This is done by iterating between two steps, the expectation step and the maximization step. The two steps described below are alternated between and the process is continued until convergence (i.e., the estimates of the parameter are almost identical; Dempster et al., 1977).

In the E- (expectation) step, the expectation of the log-likelihood function of an unknown parameter is calculated. Although the complete-data likelihood cannot be determined with a portion of the data missing, it is possible to compute the expectation of the log-likelihood function with respect to the distribution of the missing data with an initial guess about the parameter (e.g., a mean vector and a covariance matrix). In subsequent iterations, this initial guess is replaced with an estimated value from the M-step. With this information, a set of regression equations are created to predict the missing values from the observed values (Enders, 2010).

In the M- (maximization) step, the revised estimate of the parameter (e.g., a new mean vector and covariance matrix) is obtained by maximizing the expectation function

obtained in the E-step. The updated estimate of the parameter is used for the next E-step to start a new iteration (Enders, 2010).

Multiple Imputation.

In situations where a complete data set is needed multiple imputation (MI) can be used to estimate the missing values to produce a complete data set (Rubin, 1987). An MI analysis is performed using three distinct steps: the imputation phase, the analysis phase, and the pooling phase. Once the data set is imputed with values replacing the missing values, traditional statistical analysis methods can be used on the now complete data set. Similar to MLE, MI can also be used with data that are MAR, but may be biased with MNAR data (Enders, 2010).

The ultimate goal of the *imputation phase* is to generate *m* complete data sets, each containing different imputed values. This phase is divided into an imputation step and a posterior step. In the imputation step, a set of plausible values is developed. There are multiple ways to develop these values such as propensity scoring, Markov Chain Monte Carlo, and regression. For example, in the regression method the criterion in each regression equation is the variable with missing values, respectively, so that the observed values on other variables can be used to predict the missing values. The purpose of the posterior step is to vary the regression coefficients used in the imputation step, so that each data set contains different imputed values. This is done by adding a residual term to each element in both the mean vector and covariance matrix that were used in the imputation step. This creates a new set of plausible estimates from a sampling distribution (Enders, 2010).

The *analysis phase* is then used to analyze each of the complete data sets generated in the previous step, resulting in *m* statistical analyses. This step yields *m* estimates of the parameters of interest, which are then combined in the pooling phase to obtain a single estimate (Enders, 2010).

Rather than using a single parameter estimate, in the *pooling phase* all of the *m* parameter estimates obtained in the analysis phase are combined. Using the usual formula for the mean, an arithmetic average of the *m* estimates can be defined as the multiple imputation point estimate (Rubin, 1987). The resulting pooled parameter estimates not only include the finite-sample variation, but also reflect missing data uncertainty (Schafer & Graham, 2002).

Coefficient Alpha in the Presence of Missing Data.

It is important to note that when examining the effect of missing data on reliability, specifically on Cronbach's coefficient alpha as an estimate of reliability, it is assumed that "presence of missing data" is equivalent to listwise deletion. Reliability coefficients cannot be calculated on incomplete data using traditional estimation methods, hence this assumption is necessary to make.

The *missing data mechanism* of the missing values has an important effect on the estimation of coefficient alpha. While reliability estimates are expected to be unbiased at least in the MCAR case (Enders, 2004; Izquierdo & Pedrero, 2014), Enders (2003) has shown that even with MCAR data, the estimation can be biased. MAR and MNAR missing data conditions will lead to biased estimates of coefficient alpha (Enders, 2003).

Another important factor is the *amount of missing data.* A higher amount of

missing data results in a decrease in sample size. As previously discussed, a smaller sample size is associated with higher total observed score variability and consequently a smaller coefficient alpha. Therefore, coefficient alpha will underestimate reliability when a large amount of missing data is present (e.g., Izquierdo & Pedrero, 2014).

Justification for Current Study.

Despite the large amount of studies published on missing data, most of those studies are concerned with outcomes measured on a continuous scale, yet behavioral science is often concerned with item-level analyses as well (Enders, 2003). Little attention has been paid to reliability in the presence of missing data, even though it plays a ubiquitous role on measurement and applied research. Not only is there a lack of methodological research on this topic, but also applied studies often fail to report the method for handling missing data. When no indication on missing data handling is given, it is most likely to be listwise deletion, the default in most statistical software (Enders, 2004). This study aims to examine the effect of missing data on reliability under different missing data mechanisms. Furthermore, the goal is to examine whether and how this effect changes when modern missing data methods are used.

METHOD

Data Generation.

In contrast to other research focused on reliability and missing data, in this study data were simulated with specific population reliability values to begin with, in contrast to letting the reliability vary as a function of interitem correlations. Three populations with about one million observations and ten items each were drawn from a multivariate normal distribution. Since changing the number of items would have influenced the population reliability, it was held constant.

Coefficient alpha was used to create population reliability with three levels, which were achieved using a correlation matrix. The correlation matrix was composed of interitem correlations, which were chosen to create a specific reliability value. The means were set at 3.5 for half of the items, and 2.5 for the other half, to create two different levels of items. Using a depression questionnaire as an example, this would mean that participants selected higher values for the first five items and lower values for the second five items, on average. Subsequently, missing values were only inserted on items with a mean of 2.5. Each of the three populations was rounded and truncated to contain response options from one to five in order to mimic results from a common Likert-type scale. Hence, each of the three populations had about one million ordinal observations.

For each population, samples were drawn with two different sample sizes (n = 100, n = 500). Missing data was inserted into those samples based on the missing data mechanism of interest (MCAR, MAR, MNAR) and the missing data rate (p_{miss} = .05, p_{miss} = .15). For each condition, the process of sample selection, missing data insertion,

and reliability estimation was repeated 1000 times. For each of the replicated samples, coefficient alpha was calculated using listwise deletion first. Then the missing values were imputed using multiple imputation and coefficient alpha was calculated again. This resulted in 72 between-group cells with 1000 coefficient alpha values. They were analyzed by calculating bias, standardized bias, RMSE, and confidence interval coverage. R (Version 1.2.1335) was used for all data generation and analysis performed in this study.

Independent Variables.

Reliability Level.

As mentioned above, populations with different reliability levels were created using Cronbach's coefficient alpha. Reliability values of 0.7, 0.8, and 0.9 were used. This aligns with the reliability suggestions made by Langner et al. (2017), so that 0.9 represents a high reliability, 0.8 a sufficient reliability, and 0.7 a problematic reliability. The correlation matrix was determined by trial and error to result in the target reliabilities, so that interitem correlations of about .25, .38, and .60 made up the correlation matrix respectively.

Sample Size.

In this study sample size is defined as the number of observations for the total scale. For example, with a sample size of 100, each observation contains data for each of the 10 items, resulting in 1000 data points total. A decrease in sample size increases the total observed score variability in a sample and is expected to decrease coefficient alpha. To examine this effect, two different sample sizes are used. The small sample size

condition includes 100 observations and results from the goal to balance a small sample size while still being able to calculate coefficient alpha with a high level of missing data using listwise deletion. The large sample size condition includes 500 observations and aims to contain enough observations to be considered a large sample size in social sciences.

Missing Data Percentage.

Missing data were imposed on half of the items; this was considered more realistic than all items containing missing data. Specifically, missing data were imposed on the items with a mean of 2.5, so that with 10 items total there were five items with missing data and five items without missing data. The two percentages of missing data used were 5% and 15%. These percentages are based on the complete number of data points. For example, for the 5% condition with a sample size of 100, 50 data points were deleted. Divided by the number of items that were targeted for missing values, each item was missing 10 data points, which is 10% of that item. Similarly, for the 15% condition and a sample size of 100, each item had 30 missing cases and a final missing data percentage of 30%.

Missing Data Mechanism.

Missing data were generated using the MCAR, MAR, and MNAR missing data mechanisms. For the MCAR condition, random values were deleted for the first five items. Each of those items had the same amount of missing data, so that each item contained one fifth of the total missing data points.

For the MAR condition, each missing value item was paired with an item without

missing data, so that the likelihood of missingness was dependent on observed items. For example, item 1 was paired with item 6, item 2 was paired with item 7, and so on. Three different probabilities were assigned to different values on the complete item. The probability to have missing values was very high for low values on the complete item, lower for medium values, and very low for high values. This is based on the idea that participants with lower scores are more likely to have missing values. Hence, at the end each observation was paired with a certain probability to have missing values. Those probabilities were then used to delete values.

The MNAR condition was created similarly to the MAR condition. Instead of using items without missing values to create probabilities, values on the item for which values were deleted were used. For example, if an observation had a low data point on the first item, they were more likely to have a missing value on that item. The size of probabilities was kept the same as in the MAR condition.

Missing Data Techniques.

Two missing data techniques were used, listwise deletion and multiple imputation. Listwise deletion is a common traditional method to deal with missing data, while multiple imputation and maximum likelihood estimation are modern missing data handling techniques.

Listwise deletion is the default method to handle missing data in most statistical programs. However, it has been shown to only produce unbiased results under the MCAR condition. It is very difficult if not impossible to determine the missing data mechanism in a real-world scenario. In this study, listwise deletion is used under the MAR and

MNAR conditions as well, to examine the effect on reliability estimation. To use listwise deletion in the simulation, observations with at least one missing value on any item were discarded. Specifically, coefficient alpha was calculated on complete observations only.

The package "Multiple Imputation by Chained Equations" (MICE) was used for multiple imputation. The method of imputation used linear regression to predict values that fit best in place of the missing values. For each incomplete data set, 10 complete data sets were created. Even though multiple imputation theory as well as the default in R claim that five data sets are sufficient for unbiased results, Graham et al. (2007) suggest that more data sets might be needed, especially as missing data percentage increases. Therefore, 10 data sets were considered a balance between a sufficient number of data sets and minimal computational effort. The number of iterations for the estimation process of the imputed values was set at 15, as recommended by van Buuren and Groothuis-Oudshoorn (2011). For each of the 10 resulting complete data sets, coefficient alpha was calculated and pooled together, so that for each sample with missing data, there was one final alpha value combined from the complete data sets created through multiple imputation.

Maximum likelihood estimation was considered as a third missing data handling technique but was not implemented due to the lack of appropriate package to use with coefficient alpha. Even though modern missing data techniques are suggested to be straightforward to implement (Schafer & Graham 2002), there are still limitations to the practicality of their implementation.

Dependent Variables.

Standardized Bias.

Standardized bias was obtained by calculating raw bias first. When calculating raw bias, it is necessary to consider the value to which the computed coefficient alpha values are compared. In this case, the aim is to examine the deviation from the population value. Hence, bias was calculated by subtracting the appropriate population reliability value from the coefficient alpha value calculated from a sample drawn from that population. For example, when a sample was drawn from the population with reliability of 0.8, then 0.8 was subtracted from the resulting reliability estimate of that sample.

Since a certain raw bias may not be comparable across the whole metric of coefficient alpha, the standard deviation of samples with no missing data was used to compute standardized bias. Even though this is not very common, it has previously been used in missing data research (Enders, 2003). In this study, samples with the two sample sizes (N =100, 500) were drawn out of each of the three populations and their standard deviations were calculated without inserting any missing data. This resulted in six different standard deviation values based on sample size and population reliability. For each final reliability estimate, raw bias was divided by the appropriate standard deviation. For example, when a sample of 100 containing missing data was drawn from the population with reliability of 0.9, then its bias was divided by the standard deviation of a complete sample of size 100 drawn from a population with reliability of 0.9. If α_{sample} is the coefficient alpha value computed in each of the 72 between-group cells, $\alpha_{population}$ is the population reliability, and $\sigma_{complete}$ is the standard deviation from the complete sample,

then the standardized bias is

$$\text{standardized bias} = \frac{\alpha_{sample} - \alpha_{population}}{\sigma_{complete}}. \quad (3.1)$$

Root Mean Square Error.

Root Mean Square Error (RMSE) is the standard deviation of the residuals and is used to measure the difference between the estimated and the observed values. In this case, the predicted values are the population reliability values. RMSE is calculated between each of the 72 between-group cells and a vector of corresponding population reliability values. For example, for a condition in which samples were drawn out of a population with a reliability of 0.7, RMSE was calculated between the resulting vector of alpha values and a vector of the same length that contained only values of 0.7. RMSE was computed using the "hydroGOF" package in R. RMSE was calculated using the formula

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_r - P)^2}{n}}, \quad (3.2)$$

in which P is the population reliability value out of which the sample was taken, O_r is the coefficient alpha value for each of the sample replications, and n is the number of sample replications, which in this study was always 1000.

Confidence Interval Coverage.

The confidence interval for each of the alpha values was computed using the "psychometric" package in R. Two-sided intervals were created based on the number of observations (N=100, 500), the number of items (k=10), and with a confidence level of 95%. For each of the 72 between-group cells, the percentage of coefficient alpha values

that fall within the interval was calculated.

RESULTS

Reliability Level.

Standardized Bias.

The level of reliability had a noticeable impact on the standardized bias. The smallest absolute values of standardized bias occurred when reliability was 0.7. The average standardized bias was -3.02 for listwise deletion and -0.76 for multiple imputation at a reliability level of 0.7. In contrast, the highest absolute values of standardized bias occurred when reliability was 0.9. The average standardized bias was -5.29 for listwise deletion and 1.59 for multiple imputation at a reliability level of 0.9. The sign for standardized bias was always negative when listwise deletion was used, causing an underestimate of the population reliability. The lowest average bias for listwise deletion was -11.66 at a sample size of 500 and a missing data rate of 15%. For multiple imputation, the sign was negative at all reliability levels when sample size was 100 and 15% of data were missing. In addition, the sign was negative at a reliability level of 0.8 with both sample sizes at a missing data percentage of 5%. The highest average underestimate for multiple imputation was -1.79 at a reliability level of 0.7, sample size of 100, and a missing data percentage of 15%. The highest average overestimate for multiple imputation was 0.84 at a reliability level of 0.9, sample size of 500, and a missing data rate of 15%. The difference in standardized bias between listwise deletion and multiple imputation was generally large, with the difference in the largest average standardized bias being 9.87. This is partially due to missing imputation both under- and overestimating reliability under varying conditions and thereby lowering the average bias.

Root Mean Square Error.

Generally, RMSE was observed to be lower for higher reliability and higher for lower reliability for both listwise deletion and multiple imputation. These results are visible in Figures 1-4, which show the RMSE levels for each of the different conditions comparing listwise deletion to multiple imputation, different reliability levels, as well as different missing data mechanisms.

Figure 1. *RMSE when n = 100 and* p_{miss} *= 0.05*

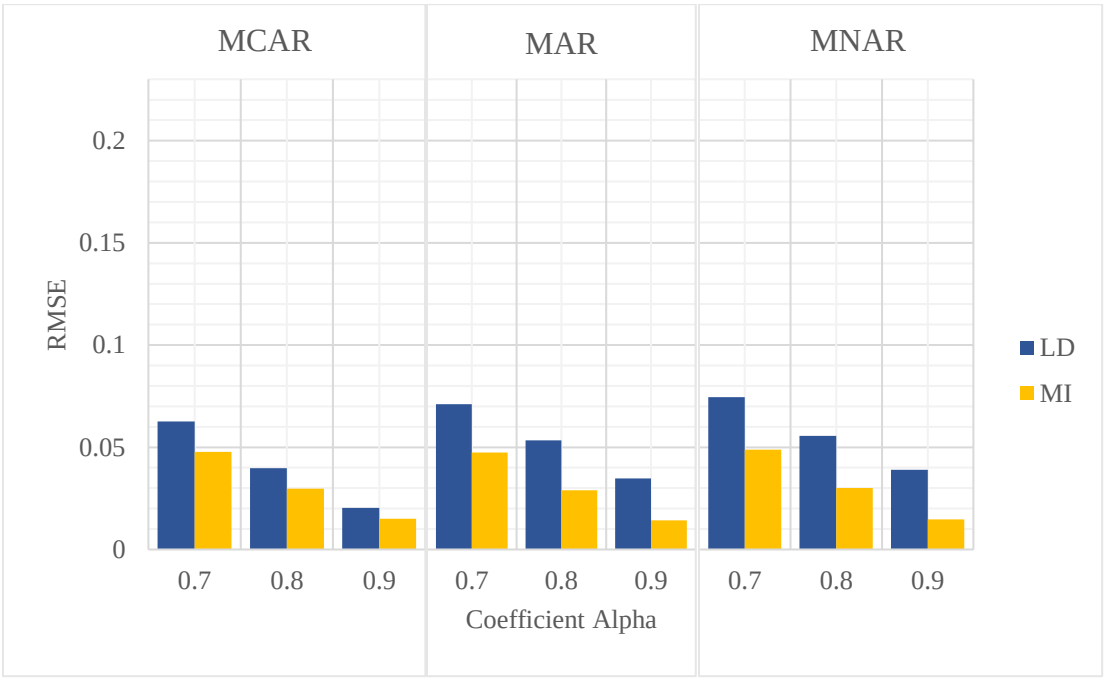


Figure 2. *RMSE when n = 100 and* p_{miss} *= 0.15*

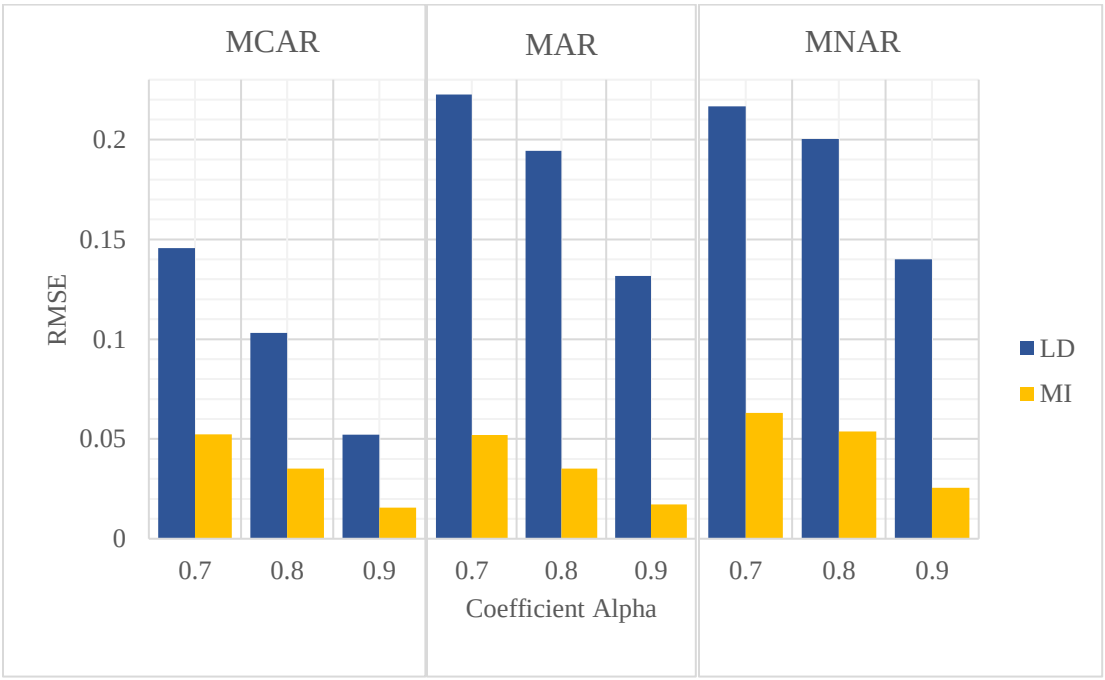


Figure 3. *RMSE when n = 500 and* p_{miss} *= 0.05*

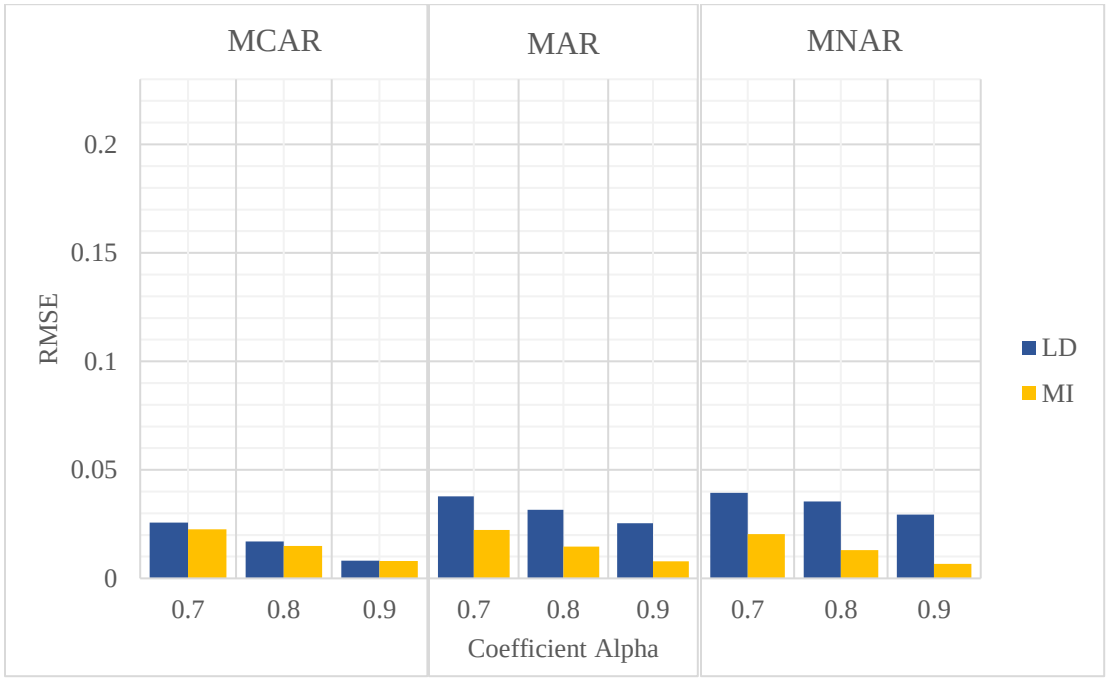
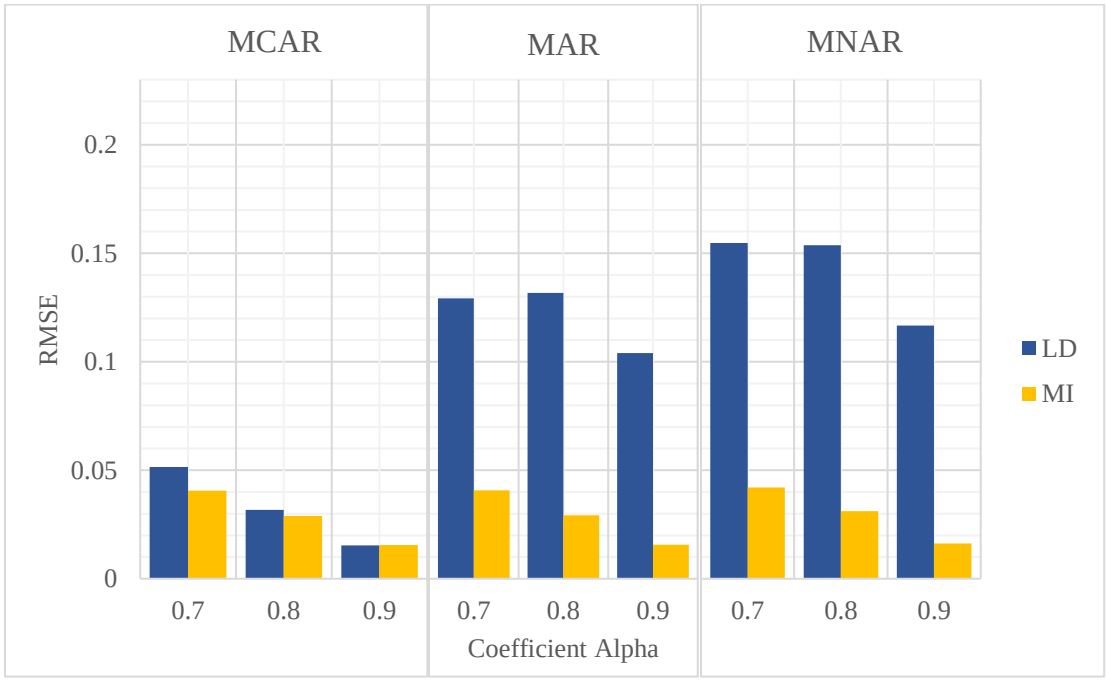


Figure 3. *RMSE when n = 500 and* p_{miss} *= 0.15*



Specifically, the highest average RMSE of 0.11 for listwise deletion occurred at a reliability level of 0.7, while the lowest RMSE of 0.06 occurred at a reliability of 0.9. The highest RMSE of 0.04 for multiple imputation was observed at a reliability of 0.7, while the lowest RMSE of 0.02 was observed at a reliability of 0.9. Lower RMSE values occurred consistently for multiple imputation as compared to listwise deletion, regardless of the reliability level. Specific ratios between RMSE values for listwise deletion and RMSE values for multiple imputation are presented in Table 1.

Table 1.

RMSE Ratios for LD and MI

Sample size	Missing data (%)	LD/MI		
		MCAR	MAR	MNAR
Coefficient alpha = 0.7				
100	5	1.31	1.50	1.52
100	15	2.79	4.29	3.43
500	5	1.14	1.70	1.94
500	15	1.27	3.17	3.68
Coefficient alpha = 0.8				
100	5	1.33	1.84	1.85
100	15	2.93	5.52	3.72
500	5	1.14	2.17	2.71
500	15	1.10	4.52	4.93
Coefficient alpha = 0.9				
100	5	1.36	2.45	2.67
100	15	3.36	7.63	5.49
500	5	1.03	3.21	4.36
500	15	0.99	6.62	7.20

Note. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; LD = listwise deletion; MI = multiple imputation.

The lowest ratio of 0.99 was observed under the MCAR condition and a reliability level of 0.9 for a sample of size 500 with 15% missing data. The highest ratio of 7.2 was observed under MNAR condition and with other factors being the same. In general, the

average ratio was higher for higher reliability, with listwise deletion showing RMSE values 2.31 times higher than listwise deletion for a reliability of 0.7 and 3.87 times higher than listwise deletion for a reliability of 0.9. The ratio difference was higher between reliabilities of 0.8 and 0.9 (1.06) than between reliabilities of 0.7 and 0.8 (0.50).

Confidence Interval Coverage.

Overall, Table 2 shows that confidence interval coverage decreased as reliability increased for both listwise deletion and multiple imputation.

Table 2.

Confidence Interval Coverage for LD and MI

		MCAR		MAR		MNAR		Total	
Sample size	Missing data (%)	LD	MI	LD	MI	LD	MI	LD	MI
		Coefficient alpha = 0.7							
100	5	0.86	0.93	0.83	0.95	0.81	0.94	0.83	0.94
100	15	0.59	0.90	0.44	0.89	0.39	0.87	0.47	0.89
500	5	0.88	0.91	0.72	0.91	0.68	0.95	0.76	0.92
500	15	0.57	0.53	0.09	0.52	0.02	0.62	0.23	0.56
		Coefficient alpha = 0.8							
100	5	0.89	0.95	0.79	0.95	0.76	0.96	0.81	0.95
100	15	0.57	0.90	0.25	0.89	0.19	0.77	0.34	0.85
500	5	0.89	0.91	0.55	0.92	0.45	0.94	0.63	0.92
500	15	0.62	0.46	0.00	0.43	0.00	0.55	0.21	0.48
		Coefficient alpha = 0.9							
100	5	0.87	0.96	0.63	0.96	0.59	0.96	0.70	0.96
100	15	0.59	0.94	0.06	0.91	0.02	0.79	0.22	0.88
500	5	0.90	0.90	0.20	0.91	0.11	0.94	0.40	0.92
500	15	0.63	0.37	0.00	0.35	0.00	0.53	0.21	0.42
		Total							
		0.74	0.81	0.38	0.80	0.34	0.82		

Note. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; LD = listwise deletion; MI = multiple imputation.

Specifically, the highest average confidence interval coverage for both listwise deletion (57%) and for multiple imputation (83%) occurred at a reliability level of 0.7. The lowest average confidence interval coverage for both listwise deletion (38%) and multiple imputation (80%) occurred at a reliability level of 0.9. The decrease in average confidence interval coverage was higher for listwise deletion (19%) than for multiple imputation (3%).

Sample Size.

Standardized Bias.

Manipulating sample size had an effect on standardized bias for both missing data techniques. The absolute value of standardized bias was smaller for the small sample size than for the large sample size for both listwise deletion (-3 compared to -4.79) and multiple imputation (-0.34 compared to 0.49). This effect was larger for listwise deletion at 1.79 than for multiple imputation at 0.15. Interestingly, the sign of the standardized bias changes between the sample sizes for multiple imputation. While for the small sample size coefficient alpha underestimated the population reliability when multiple imputation was used, it overestimated the population reliability for the large sample size, on average.

Root Mean Square Error.

The pattern for RMSE was opposite to the pattern for standardized bias for both missing data techniques. For listwise deletion, RMSE was higher when sample size was 100 (0.10) than when sample size was 500 (0.06), with a difference of 0.04. Similarly,

RMSE for multiple imputation was higher when sample size was 100 (0.03) than when sample size was 500 (0.02), with a difference of 0.01.

Confidence Interval Coverage.

Reliability estimates occurred within the confidence interval more often when sample size was small than when sample size was large. Specifically, average confidence interval coverage for listwise deletion was 56% when sample size was small and only 41% when sample size was large. Average confidence interval coverage for multiple imputation was observed to be 91% when sample size was small and 70% when sample size was large. Interestingly, the change in confidence interval coverage was larger for multiple imputation (21%) than for listwise deletion (15%).

Missing Data Percentage.

Standardized Bias.

Standardized bias was affected by the missing data percentage as well. For listwise deletion, average standardized bias increased from -1.33 at a missing data percentage of 5% to -6.47 at a missing data percentage of 15%. Conversely, for multiple imputation average standardized bias decreased from 0.13 at a missing data percentage of 5% to 0.03 at a missing data percentage of 15%. Again, this could be due to multiple imputation both under- and overestimating reliability, so that the averaging across values results in smaller average standardized bias.

Root Mean Square Error.

Overall, higher RMSE values occurred for the high missing data percentage than for the low missing data percentage for both missing data techniques. There was only a

small decrease in RMSE between the 5% missing data (0.02) and the 15% missing data (0.03) conditions when multiple imputation was used. Thereby, even the high average RMSE value for multiple imputation was lower than the low average RMSE for listwise deletion (0.04). Furthermore, the difference between the low and high average RMSE values for listwise deletion (0.09) was much higher than the difference between the low and high average RMSE values for multiple imputation (0.01). These results are displayed in Table 3.

Table 2.

RMSE for LD and MI

Sample size	Missing data (%)	MCAR		MAR		MNAR		Total	
		LD	MI	LD	MI	LD	MI	LD	MI
		Coefficient alpha = 0.7							
100	5	0.06	0.05	0.07	0.05	0.07	0.05	0.07	0.05
100	15	0.15	0.05	0.22	0.05	0.22	0.06	0.20	0.05
500	5	0.03	0.02	0.04	0.02	0.04	0.02	0.04	0.02
500	15	0.05	0.04	0.13	0.04	0.15	0.04	0.11	0.04
		Coefficient alpha = 0.8							
100	5	0.04	0.03	0.05	0.03	0.06	0.03	0.05	0.03
100	15	0.10	0.04	0.19	0.04	0.20	0.05	0.16	0.04
500	5	0.02	0.02	0.03	0.01	0.04	0.01	0.03	0.01
500	15	0.03	0.03	0.13	0.03	0.15	0.03	0.10	0.03
		Coefficient alpha = 0.9							
100	5	0.02	0.02	0.03	0.01	0.04	0.01	0.03	0.01
100	15	0.05	0.02	0.13	0.02	0.14	0.03	0.11	0.02
500	5	0.01	0.01	0.03	0.01	0.03	0.01	0.02	0.01
500	15	0.02	0.02	0.10	0.02	0.12	0.02	0.08	0.02
		Total							
		0.05	0.03	0.10	0.03	0.11	0.03		

Note. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; LD = listwise deletion; MI = multiple imputation.

Confidence Interval Coverage.

Reliability estimates were observed within the confidence interval more often when missing data percentage was small than when missing data percentage was large. Specifically, average confidence interval coverage decreased from 69% at 5% missing data to 28% at 15% missing data when listwise deletion was used. For multiple imputation, average confidence interval coverage decreased from 94% at 5% missing data to 68% at 15% missing data. Hence, the difference between the two missing data percentages was larger for listwise deletion (41%) than it was for multiple imputation (26%).

Missing Data Mechanism.

Standardized Bias.

The missing data mechanism used to add missing values had a noticeable impact on the standardized bias for both missing data techniques. For listwise deletion, standardized bias were lowest for MCAR conditions and highest for MNAR conditions (Table 4).

Table 4.

Standardized Bias for LD and MI

Sample size	Missing data (%)	MCAR		MAR		MNAR		Total	
		LD	MI	LD	MI	LD	MI	LD	MI
		Coefficient alpha = 0.7							
100	5	-0.21	0.13	-0.72	0.10	-0.83	-0.14	-0.59	0.03
100	15	-0.74	0.32	-3.31	0.44	-3.56	-0.83	-6.13	-1.79
500	5	**-0.04**	**0.56**	-1.28	0.51	-1.44	0.06	-0.92	0.38
500	15	**-0.33**	**1.79**	-5.83	1.80	-7.20	-1.72	-4.45	0.62
		Coefficient alpha = 0.8							
100	5	-0.25	0.14	-1.00	0.15	-1.10	-0.08	-0.78	-0.16
100	15	-0.87	0.27	-4.90	0.44	-5.40	-1.18	-3.72	-0.16
500	5	**-0.14**	**0.65**	-2.07	0.01	-2.45	0.07	-1.55	-0.30
500	15	**-0.24**	**2.14**	-10.13	2.15	-11.97	-2.13	-7.45	0.72
		Coefficient alpha = 0.9							
100	5	-0.20	0.20	-1.84	0.24	-2.15	-0.01	-1.40	0.14
100	15	-1.03	0.28	-8.14	0.70	-9.02	-1.19	-5.38	-0.07
500	5	**0.00**	**0.86**	-3.75	0.84	-4.44	0.33	-2.73	0.68
500	15	**-0.22**	**2.35**	-16.53	2.38	-18.67	-2.21	-11.66	0.84
		Total							
		-0.36	0.80	-4.96	0.81	-5.69	-0.75		

Note. Standardized bias values are expressed on standard error units. Situations in which listwise deletion resulted in lower standardized bias are denoted in bold. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; LD = listwise deletion; MI = multiple imputation.

The lowest standardized bias occurred at a reliability level of 0.9 with a sample size of 500 and 5% missing data (0.00), while the highest occurred at a reliability level of 0.9 with a sample size of 500 and 15% missing data (-18.67). The difference in average standardized bias was higher between MCAR and MNAR (4.6) compared to the difference between MAR and MNAR (0.73). Surprisingly, comparing average standardized bias for coefficient alpha calculated using multiple imputation under MCAR (0.80), MAR (0.81), and MNAR (-0.75), the missing data mechanism had little effect on the absolute value of the standardized bias. However, it did influence the sign. The results showed that when multiple imputation was used under MCAR and MAR conditions coefficient alpha overestimated the reliability. However, under MNAR conditions, coefficient alpha underestimated the reliability in all conditions except with sample size of 500 and 5% missing data. Another interesting finding is that under the MCAR condition for a sample size of 500, listwise deletion showed smaller bias than multiple imputation for both missing data percentages and all reliability levels.

Root Mean Square Error.

The RMSE pattern was also observed to be different between listwise deletion and multiple imputation. When listwise deletion was used, average RMSE was lowest under the MCAR condition (0.05) and highest under the MNAR condition (0.11). The difference in RMSE was higher between MCAR and MAR (0.05) than between MAR and MNAR (0.01). When multiple imputation was used, no difference between the average RMSE for the MCAR, MAR, and MNAR conditions was observed (all 0.03).

The average RMSE ratio between listwise deletion and multiple imputation was lowest for MCAR (1.65) and very similar for MAR (3.72) and MNAR (3.63).

Confidence Interval Coverage.

The confidence interval coverage between the missing data mechanisms followed the same pattern as standardized bias and RMSE. For listwise deletion, average confidence interval coverage was highest under MCAR (74%) and lowest under MNAR (34%) conditions, with the difference being higher between MCAR and MAR (36%) than between MAR and MNAR (4%). There was little difference in average confidence interval coverage for multiple imputation under the MCAR (81%), MAR (80%), and MNAR (82%) conditions.

DISCUSSION

To the best of this researcher's knowledge, there has been no research comparing listwise deletion and multiple imputation for reliability. Despite the clear superiority of modern missing data methods demonstrated in current research, such as using the EM algorithm to estimate reliability compared to traditional techniques (Enders, 2003), the implementation of these methods into practice is slow. Because multiple imputation is a potentially attractive missing data handling method, this study compared it to listwise deletion in terms of standardized bias, RMSE, and confidence interval coverage under varying conditions such as reliability level, sample size, missing data percentage, and missing data mechanism.

Reliability Level.

Overall, reliability level did affect the dependent variables. Absolute standardized bias were higher for higher levels of reliability in the case of both listwise deletion and multiple imputation, RMSE was lower for higher levels of reliability, and confidence interval coverage was lower for higher levels of reliability. This means that overall standardized bias increased as reliability increased. A possible reason for this could be that for lower reliability values, there is a comparative equal proportion of values below and above the population reliability, because it is bound between 0 and 1. For higher reliability values, there is a smaller proportion of values above, possibly making underestimation more likely than overestimation. Therefore, for lower reliability values bias are more frequently both positive and negative, resulting in lower average bias. The same pattern was found for confidence interval coverage. Overall, confidence interval

coverage decreased as reliability increased for both listwise deletion and multiple imputation. The decrease in confidence interval coverage and decrease in standardized bias for higher levels of reliability was expected as well as consistent with previous research (Enders, 2003).

Generally, the pattern for RMSE was observed to be decreasing as reliability increases for both listwise deletion and multiple imputation. In contrast to standardized bias, RMSE uses squared values and therefore the average represents the average squared distance from the predicted reliability value rather than the relative distance influenced by the sign. This could be a reason for why RMSE values behaved opposite of standardized bias values. As mentioned above, the decrease in RMSE could be due to the proportion of values above and the below a given reliability value. For example, since there are many values below and above a reliability of 0.7, bias can be the same size both in the negative and in the positive direction. However, there are fewer values above a reliability of 0.9 than for a reliability of 0.7, so that positive bias for a reliability of 0.9 can only be smaller than for the reliability of 0.7. This would result in larger bias, but in smaller RMSE.

Sample Size.

Sample size had an effect on the dependent variables for both missing data techniques as well. As expected, higher sample size resulted in lower RMSE than a smaller sample size. However, standardized bias and confidence interval coverage behaved opposite of this expectation. For a higher sample size standardized bias were higher than for a smaller sample size, on average. Similarly, confidence interval coverage

was lower for higher sample size than for lower sample size. Interestingly confidence interval coverage decreased more for multiple imputation than for listwise deletion.

The reason for the inconsistency between standardized bias, RMSE, and confidence interval coverage can have two reasons. Firstly, with a larger sample size, the denominator in the standardized bias calculation is smaller and the confidence interval is smaller as well. That way, the calculation of both standardized bias and confidence interval coverage is dependent on sample size, while the calculation of RMSE is not. Secondly, the amount of missing data is measured in percentages, so that the same missing data rate would result in more missing values total for the larger sample size than for the smaller sample size. Due to these inconsistencies it is difficult to determine whether sample size had a noticeable influence on reliability when missing data was present.

Missing Data Percentage.

As expected, standardized bias was higher, RMSE was higher, and confidence interval coverage was lower for the 15% missing data percentage than for the 5% missing data percentage for both missing data techniques. Evidently, as there are more missing data, it is more difficult to get a precise estimate of the reliability. Furthermore, an important finding is that with more missing data, the RMSE ratio between listwise deletion and multiple imputation is higher. That means that as the amount of missing data increases the more important it is to pick a missing data technique that will produce the most accurate results.

Missing Data Mechanism.

The missing data mechanism played an important role in the reliability estimates. For listwise deletion, standardized bias increased, RMSE increased, and confidence interval coverage decreased when comparing the missing data mechanism in the direction MCAR, MAR, and MNAR. For multiple imputation, the absolute standardized bias and the confidence interval coverage stayed very similar across every missing data mechanism. Interestingly, while with multiple imputation coefficient alpha overestimated reliability under the MCAR and MAR conditions, under MNAR conditions, it underestimated the reliability in almost all conditions. Since MNAR does not provide information on why data is missing (i.e. the information itself is missing), it is possible that this phenomenon is due to multiple imputation basing predictions on observed values, which are not the reason for missingness.

Interestingly, listwise deletion showed lower standardized bias than multiple imputation under MCAR conditions when sample size was 500, as denoted in bold in Table 1. Furthermore, the RMSE ratio between listwise deletion and multiple imputation is lowest under MCAR conditions and highest under MNAR conditions. This means that while for both missing data techniques RMSE increased for worse missing data mechanism conditions, this increase was much larger for listwise deletion.

Arguably, these findings most clearly show the advantage of multiple imputation compared to listwise deletion. While with multiple imputation it is possible to get results that are close to the population value (e.g. RMSE < 0.07) under all missing data

conditions, the performance of listwise deletion decreases by a lot (up to RMSE = 0.22) as soon as the missing data condition is not MCAR.

Missing Data Techniques.

Overall, multiple imputation performed better than listwise deletion under all conditions when data were missing under the MAR and MNAR mechanism. This is demonstrated by all of the dependent variables. However, under MCAR conditions, the two techniques performed very similarly, with listwise deletion even demonstrating better results in some cases. Unfortunately, it is very difficult to determine the mechanism underlying the missing data. Based on the results of this study, listwise deletion would be reasonable to use with coefficient alpha under the MCAR condition. Using Little's test of missing completely at random (Little, 1988), it is possible to determine that that mechanism is present. When both MAR and MNAR mechanisms are present, it is more reasonable to use multiple imputation with missing data.

The third missing data technique to be implemented was maximum likelihood estimation. However, to this researcher's best knowledge a package for maximum likelihood estimation with coefficient alpha is not currently available in R. This shows that even though modern missing data techniques have been shown to produce the most stable results with missing data, especially under certain conditions as discussed above, these techniques are not yet easily accessible. Especially those not experienced with research methods might find it much more difficult to implement modern missing data techniques than traditional missing data techniques such as listwise deletion.

Conclusion

Limitations.

Like all research, the present study is not without limitations. Due to a lack of resources and time constraints, it was not possible to exhaust all possible ways to deal with missing data. For example, it might be informative to examine the effect of other missing data techniques such as pairwise deletion or maximum likelihood estimation. Even though maximum likelihood estimation for coefficient alpha was not available in R, it is available in other programs. It would be particularly interesting to compare the performance of multiple imputation and maximum likelihood estimation, since those two techniques are seen as the current state of art.

Along this same line, the factors influencing reliability were limited to few categories each. This could be expanded by focusing on one specific factor while holding the others constant. For example, sample size could be gradually increased to examine its effect more closely. Similarly, the missing data percentage or reliability level could also include more categories for more detailed information. Previous research has examined the number of items as a factor (Enders, 2003), which was not included in the current study.

Another limitation of this study is that the scale of observations simulated was to be ordinal to assimilate Likert type data with specific answer categories. In this case, only ordinal values from one to five were included in the data simulation. As a result, it is not possible to make inferences beyond the conditions specified in this study.

Future Directions.

Future research will focus on how additional missing data techniques (e.g., pairwise deletion, maximum likelihood estimation) compare to listwise deletion and multiple imputation. Additionally, factors such as number of items and number of response categories will be modified. This would allow the researcher to examine their influence on coefficient alpha with missing data. Finally, given the difficulty of implementing modern missing data techniques in this study, future research will focus on making these techniques accessible to researchers without thorough methodological training. A logical procedure to include is a comparison of ways to implement different missing data techniques across statistical programs, evaluating their accessibility and utility.

Summary.

Reliability is a necessary attribute of every multiple item measure, commonly estimated by Cronbach's coefficient alpha. The current study demonstrated that coefficient alpha is sensitive to numerous factors in the presence of missing data. It is influenced by reliability level, sample size, missing data percentage, and missing data mechanism. When researchers calculate reliability in the presence of missing data, they should take these factors into account in order to evaluate possible bias in their reliability estimate.

The current study has provided evidence that when conditions are good (i.e. low missing data percentage, MCAR, large sample size), listwise deletion and multiple imputation produce very similar results. However, specifically when data are missing

under MAR and MNAR conditions, multiple imputation is advantageous. It produces fewer bias, a lower RMSE, and higher confidence interval coverage than listwise deletion. Therefore, unless researchers can be certain that their data is missing under MCAR conditions and their sample is large, they should consider using multiple imputation as their missing data technique.

Necessary to consider for practical use is also the direction of bias for the missing data techniques. When using listwise deletion, coefficient alpha is generally an underestimate of the true reliability. However, when multiple imputation is used, coefficient alpha is an overestimate when data are MCAR and MAR. However, when data are MNAR, coefficient alpha underestimates the true reliability. This is important to consider when estimating reliability in the presence of missing data.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Belmont, California: Wadsworth, Inc.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27,* 335-340.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, 22,* 302-306.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., & Meehl, P.E. (1995). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302

Cuesta I. M., & Fonseca P. E. (2014). Estimating the reliability coefficient of tests in presence of missing values. *Psicothema, 26,* 516–523.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39,* 1-38.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38,* 1006–1012.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research & Perspectives, 38*, 105-123.

Edgett, G. L. (1956). Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association, 51,* 122-131.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*, 322–337.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*, 419-436.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Ercan, I., Yazici, B., Sigirli, D., Ediz, B., & Kan, I. (2007). Examining Cronbach alpha, theta, omega reliability coefficients according to sample size. *Journal of Modern Applied Statistical Methods, 6*, 291-303.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66,* 930-944.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation. *Prevention Science, 8,* 206-213

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing, 18*, 66-67.

Hartley, J. (2013). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, *13*, 83-86.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.

Izquierdo, M., & Pedrero, F. E. (2014). Estimating the reliability coefficient of tests in presence of missing values. *Psicothema*, *26*, 516-23.

Javali, S. B., Gudaganavar, N. V., & Raj, S. M. (2011). Effect of varying sample size in estimation of coefficients of internal consistency. *WebmedCentral BIOSTATISTICS 2*:WMC001572.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*, 2276-2284.

Kuder G. F., & Richardson M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–60.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 1–55.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10-13.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Little, R. J. A., & Rubin D.B. (2002). *Statistical analysis with missing data*. 2nd ed.. New York, NY: Wiley.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493-504.

McDonald, C. J. (1999). New tools for yield improvement in integrated circuit manufacturing: Can they be applied to reliability? *Microelectronics Reliability, 39*, 731-739.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 2*, 255-273.

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

Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. *The Counseling Psychologist, 41*, 568-600.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184.

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-353.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537-560.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology, 26,* 79-85.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271-295.

Steinborn, M. B., Langner, R., Flehmig, H. C., & Huestegge, L. (2017). Methodology of performance scoring in the d2 sustained-attention test: Cumulative-reliability functions and practical guidelines. *Psychological Assessment, 30,* 339-357.

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment, 80,* 99–103.

Symonds, P. M. (1928). Factors influencing test reliability. *The Journal of Educational Psychology, 19,* 73-87.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1-67

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 98,* 22-29.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24,* 471-494.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201–223.

Yurdugül, H. (2008). Minimum sample size for Cronbach's coefficient alpha: A Monte-Carlo study. *Hacettepe University Journal of Education, 35*, 397-405.

Zhang, Z., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76*, 387–411.

APPENDIX C: R SYNTAX

```
# CREATE 3 POPULATIONS WITH DIFFERENT ALPHAS --------------------

items <- 10
kv <- c(1:items)
mu <- rep(3,items)
mu1 <- c(2.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5, 3.5)
pop.size <- 1110000

corr <- c(.251999, .380799,.598699)
populations <- list()

for (a.value1 in 1:3) {
  sigma <- matrix(corr[a.value1], items, items)
  diag(sigma) <- 1

  set.seed(13)
  data <- as.data.frame(MASS::mvrnorm(pop.size, mu, sigma))
  populations[[a.value1]] <- round(data, 0)

  num.deleted <- c(seq(-100,0),seq(6,100))
  for (column in seq_along(populations[[a.value1]])) {
    populations[[a.value1]] <- populations[[a.value1]][ ! populations[[a.value1]][, column]
%in% num.deleted, ]
} }
names(populations) <- c("pop1", "pop2", "pop3")

# SAMPLE GENERATION FOR DIFFERENT CONDITIONS ----------------------
nsamp <- 1000
m.iter <- 15
n.imp <- 10
alphaAnalysis <- numeric(n.imp)
alphacollect <- matrix(NA, nrow = nsamp, ncol = 72)

#MCAR generation for 4 conditions(repeat for the 3 populations) ------------------

alpha.ld.mcar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mcar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mcar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mcar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mcar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mcar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mcar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mcar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)
```

```
sample.size <- 100
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mcar.c1 in 1:3) {
  samp.mcar.c1 <- 1
  while (samp.mcar.c1 <= nsamp) {
    mcar.sample <- populations[[pop.mcar.c1]][sample(nrow(populations[[pop.mcar.c1]]),
sample.size), ]
    for (item.mcar.c1 in 1:5) {
      mcar.sample[, item.mcar.c1][sample(1:sample.size,n.miss)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
      alpha.ld.mcar.c1[samp.mcar.c1, pop.mcar.c1] <- psych::alpha(mcar.sample, use =
'complete.obs')$total$raw_alpha
      tempPop <- mice::mice(mcar.sample,n.imp,maxit = m.iter,method ='norm.predict',
print = FALSE)
      for (iter.mcar.c1 in 1:n.imp) {
        completed <- mice::complete(tempPop, iter.mcar.c1)
        alphaAnalysis[iter.mcar.c1] <- psych::alpha(completed)$total$raw_alpha
        alpha.mi.mcar.c1[samp.mcar.c1,pop.mcar.c1] <- mean(alphaAnalysis)
    } }, warning = function(w) {
      samp.mcar.c1 = samp.mcar.c1 - 1
      force(do.next)
    })
    samp.mcar.c1 = samp.mcar.c1 + 1
  } }

sample.size <- 100
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(25)
for (pop.mcar.c2 in 1:3) {
  samp.mcar.c2 <- 1
  while (samp.mcar.c2 <= nsamp) {
    mcar.sample <- populations[[pop.mcar.c2]][sample(nrow(populations[[pop.mcar.c2]]),
sample.size), ]
    for (item.mcar.c2 in 1:5) {
```

```
    mcar.sample[, item.mcar.c2][sample(1:sample.size,n.miss)] <- NA
   }
   delayedAssign("do.next", {next})
   result <- tryCatch({
    solve(cor(mcar.sample, use = 'complete.obs'))
    alpha.ld.mcar.c2[samp.mcar.c2, pop.mcar.c2] <- psych::alpha(mcar.sample, use =
'complete.obs')$total$raw_alpha
    tempPop <- mice::mice(mcar.sample,n.imp,maxit = m.iter,method ='norm.predict',
print = FALSE)
    for (iter.mcar.c2 in 1:n.imp) {
     completed <- mice::complete(tempPop, iter.mcar.c2)
     alphaAnalysis[iter.mcar.c2] <- psych::alpha(completed)$total$raw_alpha
     alpha.mi.mcar.c2[samp.mcar.c2,pop.mcar.c2] <- mean(alphaAnalysis)
    }}, warning = function(w) {
     samp.mcar.c2 = samp.mcar.c2 - 1
     force(do.next)
    }, error = function(e) {
     samp.mcar.c2 = samp.mcar.c2 - 1
     force(do.next)
    })
   samp.mcar.c2 = samp.mcar.c2 + 1
  } }

sample.size <- 500
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mcar.c3 in 1:3) {
  samp.mcar.c3 <- 1
  while (samp.mcar.c3 <= nsamp) {
   mcar.sample <- populations[[pop.mcar.c3]][sample(nrow(populations[[pop.mcar.c3]]),
sample.size), ]
   for (item.mcar.c3 in 1:5) {
    mcar.sample[, item.mcar.c3][sample(1:sample.size,n.miss)] <- NA
   }
   delayedAssign("do.next", {next})
   result <- tryCatch({
    alpha.ld.mcar.c3[samp.mcar.c3, pop.mcar.c3] <- psych::alpha(mcar.sample, use =
'complete.obs')$total$raw_alpha
    tempPop <- mice::mice(mcar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
    for (iter.mcar.c3 in 1:n.imp) {
```

```
      completed <- mice::complete(tempPop, iter.mcar.c3)
      alphaAnalysis[iter.mcar.c3] <- psych::alpha(completed)$total$raw_alpha
      alpha.mi.mcar.c3[samp.mcar.c3,pop.mcar.c3] <- mean(alphaAnalysis)
    }}, warning = function(w) {
    samp.mcar.c3 = samp.mcar.c3 - 1
    force(do.next)
  })
  samp.mcar.c3 = samp.mcar.c3 + 1
 } }

sample.size <- 500
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mcar.c4 in 1:3) {
 samp.mcar.c4 <- 1
 while (samp.mcar.c4 <= nsamp) {
  mcar.sample <- populations[[pop.mcar.c4]][sample(nrow(populations[[pop.mcar.c4]]),
sample.size), ]
  for (item.mcar.c4 in 1:5) {
   mcar.sample[, item.mcar.c4][sample(1:sample.size,n.miss)] <- NA
  }
  delayedAssign("do.next", {next})
  result <- tryCatch({
   alpha.ld.mcar.c4[samp.mcar.c4, pop.mcar.c4] <- psych::alpha(mcar.sample,use =
'complete.obs')$total$raw_alpha
   tempPop <- mice::mice(mcar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
   for (iter.mcar.c4 in 1:n.imp) {
    completed <- mice::complete(tempPop, iter.mcar.c4)
    alphaAnalysis[iter.mcar.c4] <- psych::alpha(completed)$total$raw_alpha
    alpha.mi.mcar.c4[samp.mcar.c4,pop.mcar.c4] <- mean(alphaAnalysis)
   }
  }, warning = function(w) {
   samp.mcar.c4 = samp.mcar.c4 - 1
   force(do.next)
  })
  samp.mcar.c4 = samp.mcar.c4 + 1
 } }

#MAR generation ----------------------------------------------
```

```
alphaAnalysis <- numeric(n.imp)
alpha.ld.mar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)

sample.size <- 100
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mar.c1 in 1:3) {
  samp.mar.c1 <- 1
  while (samp.mar.c1 <= nsamp) {
    mar.sample <- populations[[pop.mar.c1]][sample(nrow(populations[[pop.mar.c1]]),
sample.size), ]
    for (item.mar.c1 in 1:5) {
      itemplus.mar.c1 <- item.mar.c1 + 5
      p.mar <- ifelse(mar.sample[, itemplus.mar.c1] <= 2, 0.99, ifelse(itemplus.mar.c1 <=
4, 0.01, 0.05))
      mar.sample[, item.mar.c1][sample(1:sample.size,n.miss,prob = p.mar)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
      alpha.ld.mar.c1[samp.mar.c1,pop.mar.c1] <- psych::alpha(mar.sample, use =
'complete.obs')$total$raw_alpha
      tempPop <- mice::mice(mar.sample,n.imp,maxit=m.iter,method ='norm.predict', print
= FALSE)
      for (iter.mar.c1 in 1:n.imp) {
        completed <- mice::complete(tempPop, iter.mar.c1)
        alphaAnalysis[iter.mar.c1] <- psych::alpha(completed)$total$raw_alpha
        alpha.mi.mar.c1[samp.mar.c1,pop.mar.c1] <- mean(alphaAnalysis)
      }
    }, warning = function(w) {
      samp.mar.c1 = samp.mar.c1 - 1
      force(do.next)
    })
    samp.mar.c1 = samp.mar.c1 + 1
  } }
```

```
sample.size <- 100
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mar.c2 in 1:3) {
  samp.mar.c2 <- 1
  while (samp.mar.c2 <= nsamp) {
    mar.sample <- populations[[pop.mar.c2]][sample(nrow(populations[[pop.mar.c2]]),
sample.size), ]
    for (item.mar.c2 in 1:5) {
      itemplus.mar.c2 <- item.mar.c2 + 5
      p.mar <- ifelse(mar.sample[, itemplus.mar.c2] <= 2, 0.99, ifelse(itemplus.mar.c2 <=
4, 0.01, 0.05))
      mar.sample[, item.mar.c2][sample(1:sample.size,n.miss,prob = p.mar)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
      alpha.ld.mar.c2[samp.mar.c2,pop.mar.c2] <- psych::alpha(mar.sample, use =
'complete.obs')$total$raw_alpha
      tempPop <- mice::mice(mar.sample,n.imp,maxit=m.iter,method ='norm.predict', print
= FALSE)
      for (iter.mar.c2 in 1:n.imp) {
        completed <- mice::complete(tempPop, iter.mar.c2)
        alphaAnalysis[iter.mar.c2] <- psych::alpha(completed)$total$raw_alpha
        alpha.mi.mar.c2[samp.mar.c2,pop.mar.c2] <- mean(alphaAnalysis)
      }
    }, warning = function(w) {
      samp.mar.c2 = samp.mar.c2 - 1
      force(do.next)
    })
    samp.mar.c2 = samp.mar.c2 + 1
  } }

sample.size <- 500
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mar.c3 in 1:3) {
  samp.mar.c3 <- 1
```

```
  while (samp.mar.c3 <= nsamp) {
    mar.sample <- populations[[pop.mar.c3]][sample(nrow(populations[[pop.mar.c3]]),
sample.size), ]
    for (item.mar.c3 in 1:5) {
      itemplus.mar.c3 <- item.mar.c3 + 5
      p.mar <- ifelse(mar.sample[, itemplus.mar.c3] <= 2, 0.99, ifelse(itemplus.mar.c3 <=
4, 0.01, 0.05))
      mar.sample[, item.mar.c3][sample(1:sample.size,n.miss,prob = p.mar)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
      alpha.ld.mar.c3[samp.mar.c3,pop.mar.c3] <- psych::alpha(mar.sample, use =
'complete.obs')$total$raw_alpha
      tempPop <- mice::mice(mar.sample,n.imp,maxit=m.iter,method ='norm.predict', print
= FALSE)
      for (iter.mar.c3 in 1:n.imp) {
        completed <- mice::complete(tempPop, iter.mar.c3)
        alphaAnalysis[iter.mar.c3] <- psych::alpha(completed)$total$raw_alpha
        alpha.mi.mar.c3[samp.mar.c3,pop.mar.c3] <- mean(alphaAnalysis)
      }
    }, warning = function(w) {
      samp.mar.c3 = samp.mar.c3 - 1
      force(do.next)
    })
    samp.mar.c3 = samp.mar.c3 + 1
  } }

sample.size <- 500
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mar.c4 in 1:3) {
  samp.mar.c4 <- 1
  while (samp.mar.c4 <= nsamp) {
    mar.sample <- populations[[pop.mar.c4]][sample(nrow(populations[[pop.mar.c4]]),
sample.size), ]
    for (item.mar.c4 in 1:5) {
      itemplus.mar.c4 <- item.mar.c4 + 5
      p.mar <- ifelse(mar.sample[, itemplus.mar.c4] <= 2, 0.99, ifelse(itemplus.mar.c4 <=
4, 0.01, 0.05))
      mar.sample[, item.mar.c4][sample(1:sample.size,n.miss,prob = p.mar)] <- NA
    }
```

```
   delayedAssign("do.next", {next})
   result <- tryCatch({
     alpha.ld.mar.c4[samp.mar.c4,pop.mar.c4] <- psych::alpha(mar.sample, use =
'complete.obs')$total$raw_alpha
     tempPop <- mice::mice(mar.sample,n.imp,maxit=m.iter,method ='norm.predict', print
= FALSE)
     for (iter.mar.c4 in 1:n.imp) {
       completed <- mice::complete(tempPop, iter.mar.c4)
       alphaAnalysis[iter.mar.c4] <- psych::alpha(completed)$total$raw_alpha
       alpha.mi.mar.c4[samp.mar.c4,pop.mar.c4] <- mean(alphaAnalysis)
     }
   }, warning = function(w) {
     samp.mar.c4 = samp.mar.c4 - 1
     force(do.next)
   })
   samp.mar.c4 = samp.mar.c4 + 1
 } }

#MNAR generation ------------------------------------------------------------------------
alphaAnalysis <- numeric(n.imp)
alpha.ld.mnar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mnar.c1 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mnar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mnar.c2 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mnar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mnar.c3 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.ld.mnar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)
alpha.mi.mnar.c4 <- matrix(NA, nrow = nsamp, ncol = 3)

sample.size <- 100
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mnar.c1 in 1:3) {
 samp.mnar.c1 <- 1
 while (samp.mnar.c1 <= nsamp) {
   mnar.sample <- populations[[pop.mnar.c1]][sample(nrow(populations[[pop.mnar.c1]]),
sample.size), ]

   for (item.mnar.c1 in 1:5) {
     p.mnar <- ifelse(mnar.sample[, item.mnar.c1] <= 2, 0.99, ifelse(item.mnar.c1 <= 4,
0.01, 0.05))
```

```
    mnar.sample[, item.mnar.c1][sample(1:sample.size,n.miss,prob = p.mnar)] <- NA
   }
   delayedAssign("do.next", {next})
   result <- tryCatch({
    alpha.ld.mnar.c1[samp.mnar.c1,pop.mnar.c1] <- psych::alpha(mnar.sample, use =
'complete.obs')$total$raw_alpha
    tempPop <- mice::mice(mnar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
    for (iter.mnar.c1 in 1:n.imp) {
     completed <- mice::complete(tempPop, iter.mnar.c1)
     alphaAnalysis[iter.mnar.c1] <- psych::alpha(completed)$total$raw_alpha
     alpha.mi.mnar.c1[samp.mnar.c1,pop.mnar.c1] <- mean(alphaAnalysis)
    }
   }, warning = function(w) {
    samp.mnar.c1 = samp.mnar.c1 - 1
    force(do.next)
   })
   samp.mnar.c1 = samp.mnar.c1 + 1
  } }

sample.size <- 100
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mnar.c2 in 1:3) {
  samp.mnar.c2 <- 1
  while (samp.mnar.c2 <= nsamp) {
   mnar.sample <- populations[[pop.mnar.c2]][sample(nrow(populations[[pop.mnar.c2]]),
sample.size), ]
   for (item.mnar.c2 in 1:5) {
    p.mnar <- ifelse(mnar.sample[, item.mnar.c2] <= 2, 0.99, ifelse(item.mnar.c2 <= 4,
0.01, 0.05))
    mnar.sample[, item.mnar.c2][sample(1:sample.size,n.miss,prob = p.mnar)] <- NA
   }
   delayedAssign("do.next", {next})
   result <- tryCatch({
    alpha.ld.mnar.c2[samp.mnar.c2,pop.mnar.c2] <- psych::alpha(mnar.sample, use =
'complete.obs')$total$raw_alpha
    tempPop <- mice::mice(mnar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
    for (iter.mnar.c2 in 1:n.imp) {
     completed <- mice::complete(tempPop, iter.mnar.c2)
```

```
      alphaAnalysis[iter.mnar.c2] <- psych::alpha(completed)$total$raw_alpha
      alpha.mi.mnar.c2[samp.mnar.c2,pop.mnar.c2] <- mean(alphaAnalysis)
     }
    }, warning = function(w) {
     samp.mnar.c2 = samp.mnar.c2 - 1
     force(do.next)
    })
    samp.mnar.c2 = samp.mnar.c2 + 1
  } }

sample.size <- 500
missp <- .05
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mnar.c3 in 1:3) {
  samp.mnar.c3 <- 1
  while (samp.mnar.c3 <= nsamp) {
    mnar.sample <- populations[[pop.mnar.c3]][sample(nrow(populations[[pop.mnar.c3]]),
sample.size), ]
    for (item.mnar.c3 in 1:5) {
     p.mnar <- ifelse(mnar.sample[, item.mnar.c3] <= 2, 0.99, ifelse(item.mnar.c3 <= 4,
0.01, 0.05))
     mnar.sample[, item.mnar.c3][sample(1:sample.size,n.miss,prob = p.mnar)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
     alpha.ld.mnar.c3[samp.mnar.c3,pop.mnar.c3] <- psych::alpha(mnar.sample, use =
'complete.obs')$total$raw_alpha
     tempPop <- mice::mice(mnar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
     for (iter.mnar.c3 in 1:n.imp) {
      completed <- mice::complete(tempPop, iter.mnar.c3)
      alphaAnalysis[iter.mnar.c3] <- psych::alpha(completed)$total$raw_alpha
      alpha.mi.mnar.c3[samp.mnar.c3,pop.mnar.c3] <- mean(alphaAnalysis)
     }
    }, warning = function(w) {
     samp.mnar.c3 = samp.mnar.c3 - 1
     force(do.next)
    })
    samp.mnar.c3 = samp.mnar.c3 + 1
  } }
```

```
sample.size <- 500
missp <- .15
missn <- missp * (items * sample.size)
n.miss <- missn/(items/2)

set.seed(20)
for (pop.mnar.c4 in 1:3) {
  samp.mnar.c4 <- 1
  while (samp.mnar.c4 <= nsamp) {
    mnar.sample <- populations[[pop.mnar.c4]][sample(nrow(populations[[pop.mnar.c4]]),
sample.size), ]
    for (item.mnar.c4 in 1:5) {
      p.mnar <- ifelse(mnar.sample[, item.mnar.c4] <= 2, 0.99, ifelse(item.mnar.c4 <= 4,
0.01, 0.05))
      mnar.sample[, item.mnar.c4][sample(1:sample.size,n.miss,prob = p.mnar)] <- NA
    }
    delayedAssign("do.next", {next})
    result <- tryCatch({
      alpha.ld.mnar.c4[samp.mnar.c4,pop.mnar.c4] <- psych::alpha(mnar.sample, use =
'complete.obs')$total$raw_alpha
      tempPop <- mice::mice(mnar.sample,n.imp,maxit=m.iter,method ='norm.predict',
print = FALSE)
      for (iter.mnar.c4 in 1:n.imp) {
        completed <- mice::complete(tempPop, iter.mnar.c4)
        alphaAnalysis[iter.mnar.c4] <- psych::alpha(completed)$total$raw_alpha
        alpha.mi.mnar.c4[samp.mnar.c4,pop.mnar.c4] <- mean(alphaAnalysis)
      }
    }, warning = function(w) {
      samp.mnar.c4 = samp.mnar.c4 - 1
      force(do.next)
    })
    samp.mnar.c4 = samp.mnar.c4 + 1
  } }

alphacollect <- data.frame(
  alpha.ld.mcar.c1, alpha.ld.mcar.c2, alpha.ld.mcar.c3, alpha.ld.mcar.c4,
  alpha.mi.mcar.c1, alpha.mi.mcar.c2, alpha.mi.mcar.c3, alpha.mi.mcar.c4,
  alpha.ld.mar.c1,  alpha.ld.mar.c2,  alpha.ld.mar.c3,  alpha.ld.mar.c4,
  alpha.mi.mar.c1,  alpha.mi.mar.c2,  alpha.mi.mar.c3,  alpha.mi.mar.c4,
  alpha.ld.mnar.c1, alpha.ld.mnar.c2, alpha.ld.mnar.c3, alpha.ld.mnar.c4,
  alpha.mi.mnar.c1, alpha.mi.mnar.c2, alpha.mi.mnar.c3, alpha.mi.mnar.c4
  )
labels <- c(
  "a1-MCAR-ld-c1", "a2-MCAR-ld-c1", "a3-MCAR-ld-c1",
```

```
  "a1-MCAR-ld-c2", "a2-MCAR-ld-c2", "a3-MCAR-ld-c2",
  "a1-MCAR-ld-c3", "a2-MCAR-ld-c3", "a3-MCAR-ld-c3",
  "a1-MCAR-ld-c4", "a2-MCAR-ld-c4", "a3-MCAR-ld-c4",

  "a1-MCAR-mi-c1", "a2-MCAR-mi-c1", "a3-MCAR-mi-c1",
  "a1-MCAR-mi-c2", "a2-MCAR-mi-c2", "a3-MCAR-mi-c2",
  "a1-MCAR-mi-c3", "a2-MCAR-mi-c3", "a3-MCAR-mi-c3",
  "a1-MCAR-mi-c4", "a2-MCAR-mi-c4", "a3-MCAR-mi-c4",

  "a1-MAR-ld-c1", "a2-MAR-ld-c1", "a3-MAR-ld-c1",
  "a1-MAR-ld-c2", "a2-MAR-ld-c2", "a3-MAR-ld-c2",
  "a1-MAR-ld-c3", "a2-MAR-ld-c3", "a3-MAR-ld-c3",
  "a1-MAR-ld-c4", "a2-MAR-ld-c4", "a3-MAR-ld-c4",

  "a1-MAR-mi-c1", "a2-MAR-mi-c1", "a3-MAR-mi-c1",
  "a1-MAR-mi-c2", "a2-MAR-mi-c2", "a3-MAR-mi-c2",
  "a1-MAR-mi-c3", "a2-MAR-mi-c3", "a3-MAR-mi-c3",
  "a1-MAR-mi-c4", "a2-MAR-mi-c4", "a3-MAR-mi-c4",

  "a1-MNAR-ld-c1", "a2-MNAR-ld-c1", "a3-MNAR-ld-c1",
  "a1-MNAR-ld-c2", "a2-MNAR-ld-c2", "a3-MNAR-ld-c2",
  "a1-MNAR-ld-c3", "a2-MNAR-ld-c3", "a3-MNAR-ld-c3",
  "a1-MNAR-ld-c4", "a2-MNAR-ld-c4", "a3-MNAR-ld-c4",

  "a1-MNAR-mi-c1", "a2-MNAR-mi-c1", "a3-MNAR-mi-c1",
  "a1-MNAR-mi-c2", "a2-MNAR-mi-c2", "a3-MNAR-mi-c2",
  "a1-MNAR-mi-c3", "a2-MNAR-mi-c3", "a3-MNAR-mi-c3",
  "a1-MNAR-mi-c4", "a2-MNAR-mi-c4", "a3-MNAR-mi-c4"
  )

colnames(alphacollect) <- labels

#CONTROL condition (no missing data) ---------------------------------------------

alpha.con1 <- data.frame()
alpha.con2 <- data.frame()
alpha.con <- data.frame()

sample.size <- 100
set.seed(20)
for (pop.con in 1:3) {
  for (samp.con in 1:nsamp) {
    con.sample <- populations[[pop.con]][sample(nrow(populations[[pop.con]]),
sample.size), ]
```

```
   alpha.con1[samp.con,pop.con] <- psych::alpha(con.sample)$total$raw_alpha
  } }

sample.size <- 500
set.seed(20)
for (pop.con in 1:3) {
  for (samp.con in 1:nsamp) {
    con.sample <- populations[[pop.con]][sample(nrow(populations[[pop.con]]),
sample.size), ]
    alpha.con2[samp.con,pop.con] <- psych::alpha(con.sample)$total$raw_alpha
  } }

alpha.con <- data.frame(alpha.con1, alpha.con2)

#ANALYSIS OF THE FINAL DATA SET OF ALPHAS ----------------------------

depvars <- data.frame(matrix(NA, nrow = 72, ncol = 4))
rownames(depvars) <- labels
colnames(depvars) <- c("raw bias", "std bias", "RMSE", "CI coverage")

# raw and standardized bias
alphameans <- (rep(c(.7,.8,.9), length.out = 72))
column.means <- colMeans(alphacollect)
for (obs.rawbias in 1:72) {
 depvars[obs.rawbias, 1] <- column.means[obs.rawbias] - alphameans[obs.rawbias]
}
sd.unordered <- apply(alpha.con, 2, sd)
sd.ordered <- rep(c(rep(sd.unordered[1:3], times = 2), rep(sd.unordered[4:6], times = 2)),
length.out = 72)
for (obs.stdbias in 1:72) {
  depvars[obs.stdbias, 2] <- depvars[obs.stdbias, 1] / sd.ordered[obs.stdbias]
}

# RMSE
alpha.predict <- data.frame(matrix(rep(c(.7,.8,.9), each = nsamp, times = 24), ncol = 72,
byrow = FALSE))
for (obs.rmse in 1:72) {
  depvars[obs.rmse, 3] <- hydroGOF::rmse(alphacollect[ ,obs.rmse], alpha.predict[
,obs.rmse])
}

# Confidence interval coverage
conf.inter <- data.frame(matrix(NA, nrow = 3, ncol = 2))
number <- c(.7, .8, .9)
```

```
for (obs.conf in 1:3) {
  conf.inter[obs.conf,1] <- psychometric::alpha.CI(number[obs.conf], 10, 100, level =
0.95, onesided = FALSE)$LCL
  conf.inter[obs.conf,2] <- psychometric::alpha.CI(number[obs.conf], 10, 100, level =
0.95, onesided = FALSE)$UCL
}
for (obs.conf in 4:6) {
  conf.inter[obs.conf,1] <- psychometric::alpha.CI(number[obs.conf - 3], 10, 500, level =
0.95, onesided = FALSE)$LCL
  conf.inter[obs.conf,2] <- psychometric::alpha.CI(number[obs.conf - 3], 10, 500, level =
0.95, onesided = FALSE)$UCL
}

fun <- numeric(length = nsamp)
conf.predict <- rep(c(1,2,3,1,2,3,4,5,6,4,5,6), times = 6)
for (obs.conftf in 1:72) {
  for (sep.alpha in 1:nsamp) {
   if (alphacollect[sep.alpha,obs.conftf] > conf.inter[conf.predict[obs.conftf],1] &&
alphacollect[sep.alpha,obs.conftf] < conf.inter[conf.predict[obs.conftf],2]) {
     fun[sep.alpha] <- 1
   } else {fun[sep.alpha] <- 0}
  }
  depvars[obs.conftf,4] <- sum(fun)/nsamp
}
```