JOURNAL OF APPLIED MEASUREMENT, *20*(2), 154-166

Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation

Glenn Thomas Waterbury
James Madison University

This simulation study explores the effects of missing data mechanisms, proportions of missing data, sample size, and test length on the biases and standard errors of item parameters using the Rasch measurement model. When responses were missing completely at random (MCAR) or missing at random (MAR), item parameters were unbiased. When responses were missing not at random (MNAR), item parameters were severely biased, especially when the proportion of missing responses was high. Standard errors were primarily affected by sample size, with larger samples associated with smaller standard errors. Standard errors were inflated in MCAR and MAR conditions, while MNAR standard errors were similar to what they would have been, had the data been complete. This paper supports the conclusion that the Rasch model can handle varying amounts of missing data, provided that the missing responses are not MNAR.

Requests for reprints should be sent to Tom Waterbury, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg, VA 22807, USA, e-mail: waterbgt@jmu.edu.

Introduction

According to prominent Rasch measurement model proponent Benjamin Wright, "…a useful measurement model for constructing inference from observation must be unaffected by missing data" (Wright and Mok, p. 3, 2004). Missing data is not uncommon in educational and psychological testing. When Ben Wright stated that a measurement model must be "unaffected" by missing data, he could have also phrased it as "a useful measurement model can ignore missing data." If missing data is ignorable, then parameters from that model will be unaffected by the presence of missing data. To determine whether missing data can be ignorable, many things must be taken into consideration. For example, the mechanism underlying the missing data, the model being specified, and the estimation technique all play a crucial role in determining the ignorability of missing data. The purpose of this simulation study will be to examine the impact of different missing data mechanisms on the biases and standard errors of item parameter estimates of the Rasch model, specifically using simulated data and Winsteps (Linacre, 2015) for parameter estimation.

Before discussing the Rasch model specifically, it is important to differentiate between types of missing data. Not all missing data are created equal. In his seminal article, Rubin (1976) defined three mechanisms underlying missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Data may be considered MCAR when missingness on a variable is unrelated to the values of the variable itself and to any other measured variable. Essentially, the observed data are simply a random sample of the complete data. Using a random number generator to delete half the scores on a certain variable would constitute missingness that is MCAR.

Data are considered MAR when missingness on a variable is related to other measured variables, but not to values of the variable with missingness itself. For example, imagine a data set with two variables: height and basketball ability. The basketball ability variable contains missing data. Specifically, imagine that people under a certain height did not have their basketball ability tested. In this case, the missing data would be considered MAR because missingness on basketball ability was completely determined by height.

Data are considered MNAR when missingness on a variable is related to values of the variable itself. For example, imagine a variable that contains the annual salary for everyone in a data set. This variable contains missing data. Specifically, those with very high and very low income were uncomfortable reporting their income and thus did not report it. This would be an example of data that are MNAR, as the missingness is related to the values that *would* have been reported had the data been complete.

Many techniques exist for the handling of missing data. Some techniques are much more effective than others. For example, the most common traditional technique for handling missing data has been listwise deletion (Peugh and Enders, 2004). Listwise deletion entails deleting all cases in a dataset for which there is incomplete data. The appeal of listwise deletion is its simplicity, as the deletion of any case with missing data essentially falsely creates a new dataset with complete data.

One state of the art technique for accommodating missing data is maximum likelihood estimation. See Enders (2005) for an excellent description of maximum likelihood estimation. Briefly, the purpose of maximum likelihood estimation is to identify parameters that maximize the likelihood of the data. When data are missing, cases with missing data are not deleted. Instead, all available information from each case are used in the likelihood function that calculates the likelihood of a set of parameters, given the data. When a data point from a certain case on a certain variable is missing, that missing data point is simply not included in the likelihood function.

There is plentiful literature regarding the ability of maximum likelihood to handle missing data (Arbuckle, 1996; Enders and Bandalos, 2001; Enders, 2010). When data are MCAR or

MAR, parameter estimates from maximum likelihood estimation should be unbiased. This stands in contrast to the popular technique of listwise deletion, which yields biased parameters when data are MAR.

Most studies regarding missing data and maximum likelihood estimation have been conducted with continuous variables. With educational and psychological testing, data are usually categorical, and often dichotomous. In many testing situations, item response theory or Rasch measurement models are used to analyze responses to dichotomously scored test items. Both IRT and Rasch models posit that the probability of a correct response on an item is a function of the examinee's ability and one or more item parameters. In the case of Rasch models, the item's difficulty is the sole item parameter. When item responses are missing when using IRT or Rasch models, there is a fundamental decision to make: consider items with missing responses as incorrect, or treat them as missing data. Research has shown that in IRT, scoring missing data as incorrect is generally the worst of the two decisions, in terms of biasing parameter estimates (Lord, 1974; Mislevy and Wu, 1988; De Ayala, Plake, and Impara, 2001). Instead, a better solution is to treat missing data as missing and accommodate the missingness in some way.

Very little research exists explicitly concerning missing data and the Rasch model. Linacre (2004; 2015) has stated that many Rasch software packages can readily handle missing data, including the popularly used program Winsteps. However, detail is lacking in terms of the *types* of missing data that can be handled by Winsteps or other Rasch software. Andrich and co-authors (2012; 2014; 2016) have recently formalized a method for adjusting item and person parameters that are biased by the presence of correct guessing. With this method, all person/item encounters with a probability of a correct response below a certain probability threshold are converted to *missing data* (this process can be done in Winsteps using the cut-low command). This is done regardless of whether the response to the item was correct or incorrect. The logic is that the person/item encounters being converted to missing data constitute responses that were likely to have been guessed. Andrich and co-authors, using the software package RUMM (Andrich, Sheridan, and Luo, 2013), found that this guessing adjustment strategy all-but eliminated bias in item and person parameters. In a simulation study using Winsteps, Waterbury and DeMars (2019) found that the guessing adjustment did not eliminate bias, but reduced it substantially.

While not mentioned formally in studies investigating this method, the guessing adjustment induces data that are MAR. While data were not missing completely at random, missingness on each item depended entirely on variables in the model: the item's difficulty and the person's ability. That the guessing adjustment creates data that are MAR, as opposed to MNAR, is the reason it is a tenable strategy for reducing the impact of correct guessing. While these studies lend credence to the fact that several Rasch measurement software can accommodate certain types of missing data, missingness was not a focus of the studies. Nor could the effects of missingness be completely isolated from the effect of correct guessing in the data. Therefore, the purpose of this study is to examine the effects of sample size, test length, proportion of missing responses, and missing data mechanism on the biases and standard errors of Rasch item parameters using Winsteps.

Method

Data were simulated using SAS 9.4. Data were simulated using the dichotomous Rasch model (Rasch, 1960), the equation for which is shown below:

$$\Pr\{x_{ni}=1\}\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}, \quad (1)$$

where x_{ni} is the observed response (0 or 1) from person n on item i, β_n is the ability of person n and δ_i is the difficulty of item i. Unlike more complex IRT models, the Rasch model requires that the probability of a correct response be a function solely of the difficulty of the item and the ability of the person. This is a strict requirement. Yet,

when this requirement is met, the Rasch model yields desirable properties of objective measurement and sufficiency of total scores for estimation of parameters.

Data were manipulated in the following ways: a) sample size, b) test length, c) proportion of missing responses on items with missingness, and d) the type of missing data mechanism. This yielded a fully crossed design of sample size x test length x proportion of missing responses x missing data mechanism = 3 x 2 x 2 x 3 = 36 conditions. Each factor of the study is described in more detail below, along with the criteria for comparison across conditions: bias and standard errors.

Sample size was the first factor of the study. This factor had three levels: 500, 1,000, or 2,000 examinees. Examinees were always drawn from a standard normal distribution, with a mean ability of zero and a variance of one. It was not anticipated that sample size would affect the bias of item parameters. It, of course, would affect the standard errors of item parameters, with larger sample sizes yielding smaller standard errors. Of interest were possible interactions between sample size and missing data mechanism on standard errors, or interactions between sample size and proportion of missing data. Perhaps the effect of sample size on the standard errors of item parameters depended on the type or proportion of missingness.

The second factor was test length. The test consisted of either 60 or 30 items. Items were always uniformly distributed, with a difficulty origin of zero. When test length was 60 items, item difficulties ranged from –3 to 3, in increments of .1 logits. When test length was 30 items, item difficulties still ranged from –3 to 3, but in increments of .2 logits. Regardless of test length, items were arranged in order of difficulty, with the first item being the easiest and the last item being the hardest. When test length was 60, items 50, 52, 54, 56, 58, and 60 were the items that contained missing data. When test length was 30, items 26, 28, and 30 contained missing data. Therefore, irrespective of test length, the percentage of items with missing data was constrained to 10%. Given the location of the items with missing data on the test, items with missing data were the most difficult items. Reasons for this will be explained subsequently.

The third factor of the study was the proportion of missing responses on items with missingness. This factor had two levels: .2 or .5. Theoretically, when data are MCAR or MAR, item parameters should be unbiased, regardless of the amount of missing data. However, when data are MNAR, this invariance of bias across proportions of missing data may not hold. Additionally, the proportion of missing data should have an effect on the standard errors of item parameters. When data are missing, parameters are estimated using less information than if data were complete, thus decreasing the precision of estimates and increasing the standard errors.

The final factor, and the crux of this study, was the missing data mechanism underlying the missing item responses. This factor had three levels: MCAR, MAR, and MNAR. Missing data mechanisms were completely separate. There were no conditions where an item contained missing responses generated by multiple missing data mechanisms, such as both MAR and MNAR. This was done intentionally, in order to isolate the effects of each missing data mechanism.

MCAR conditions were created by randomly deleting 20% or 50% of responses for the items with missingness. In these conditions, missing responses did not depend on the values of the missing responses themselves, nor any other items on the test. It was as if the observed responses to the items were simply a random sample of the complete set of responses. Item parameters should not be biased in the MCAR conditions, regardless of the levels of the other factors.

MAR conditions were created by basing missing responses entirely on the response to the previous item. For example, recall that the 60 item test contained missing responses on items 50, 52, 54, 56, 58, and 60. For examinees that answered item 49 *correctly*, they were *guaranteed* to have a response for item 50. If an examinee

answered *incorrectly* on item 49, that examinee had a probability of a missing response on item 50. The probability of a missing response for an item was always chosen so that either 20% or 50% of responses were missing on that item, depending on the level of the proportion of missingness factor. For example, with a true difficulty of 2 logits, approximately 87% of examinees from a standard normal distribution were expected to answer item 49 incorrectly. Subsequently, in order to have 50% of responses to item 50 be missing, the probability of missingness on item 50 for examinees who were incorrect on item 49 was approximately .575. The probability had to be greater than .5 because the probability of a missing response applied only to examinees who were *incorrect* on item 49, not the *entire examinee sample*. Because each item had a different true difficulty, each item with missing responses had its own unique probability of missingness. However, the proportion of missing responses on each item with missingness always amounted to .2 or .5, depending on the condition.

This pattern of missing data constitutes MAR because missing responses depended solely on another item in the data set. This type of MAR mechanism was chosen to mimic a plausible cause of missingness in educational testing. Imagine items 49 and 50 are based on very similar knowledge or procedures. Examinees who score incorrectly on item 49 may feel compelled to skip the next item because they feel they cannot answer it correctly, given the similarity of content and difficulty of the two items. This type of missingness seems more plausible with difficult items, as opposed to easy items. For an easy item, less students will answer the item incorrectly and thus have less reason to skip the subsequent item. This MAR mechanism was simulated for all items with missing responses in the MAR conditions. As with MCAR conditions, item parameters in MAR conditions should not be biased.

MNAR conditions were created by basing missing responses entirely on the item responses of the items with missingness. A complete data set was first simulated, so that every item/examinee encounter contained a response. Then, any examinee with an incorrect response had a probability of their response being converted to missing. For example, if an examinee answered item 50 incorrectly, that examinee would have a probability of his/her response being converted to missing data. As with MAR conditions, the exact probability of missingness was chosen so that the proportion of missing responses was either .2 or .5. This constitutes MNAR, as the missing responses depended on the values of the item responses themselves, had the responses actually been given. This represents a plausible example of missing data in educational testing. An examinee encounters an item that is much too difficult. Out of frustration or the realization that his/her response would be incorrect, the examinee instead skips the item. This type of MNAR data could be created because this is a simulation study. In practice, there is no way to discern whether data are MNAR. When missing item responses were MNAR, it was expected that item parameters would be biased, especially when the proportion of missing data was .5.

All Rasch analyses were conducted using the Winsteps computer program V. 3.92 (Linacre, 2015). Winsteps uses joint maximum likelihood estimation to obtain parameter estimates, which relies on successive iterations to find the optimal estimates. Unlike traditional JMLE, which relies on the Newton-Raphson procedure, Winsteps uses a proportional curve fitting algorithm. Given the importance of estimation in regards to missing data, a brief overview of the Winsteps estimation process is warranted.

Briefly, the proportional curve fitting algorithm relies on the fact that the relation between Rasch person measures and person total scores is necessarily monotonic and takes the form of a logistic ogive. This is known as a test characteristic curve (TCC). While discussed much less, the same monotonic relation, modeled by a logistic ogive, is true of Rasch item measures and total scores for items. This could be called a person characteristic curve (PCC). The proportional curve fitting algorithm begins with starting estimates for every person ability and item difficulty, usually arrived upon by using the normal

approximation (PROX) algorithm (Wright and Masters, 1982). For each person, the initial ability estimate and a slightly different estimate are used to estimate the slope and intercept of a hypothetical TCC for each person. Once the slope and intercept are estimated, a TCC is constructed for each person. Using the observed raw total score for each person, the corresponding Rasch ability is identified on the TCC and that becomes the new person ability estimate. This new estimate, and a nearby estimate, are then used to estimate a new slope and intercept to construct a new hypothetical TCC. That new TCC is then used to find a new ability estimate corresponding to the observed raw score. This process is iterated multiple times until the maximum change of any ability estimate is within a specified convergence criteria, often .001 or .0001 logits (the default convergence criteria for Winsteps is .0001 logits). It is known as proportional curve fitting because the algorithm iterates until the slope and intercept of the TCC constructed for each person most closely fits the Rasch model implied TCC.

The estimation of item difficulties occurs simultaneously with the estimation of abilities and follows an identical algorithm. The only difference is that hypothetical *person* characteristic curves are constructed for each item, instead of *test* characteristic curves for each person. While computationally distinct from JMLE with the Newton-Raphson procedure, proportional curve fitting yields functionally equivalent results. For more specific details about the proportional curve fitting algorithm itself, see Meyer and Hailey (2012) for an excellent description.

In regards to missing data, missing responses with proportional curve fitting are treated the same as is done using traditional maximum likelihood. With traditional maximum likelihood estimation, missing data is ignored and the likelihood function includes all available information. Proportional curve fitting follows the same approach. For estimating person abilities with the curve fitting algorithm, raw total person scores and TCC's are computed using only the items a person responded too. Similarly, when estimating item difficulties, raw total item scores and PCC's are computed using only the people from whom an item had a response. This ignoring of missing data should be tenable when item responses are MCAR or MAR, but not when responses are MNAR.

The two criteria for comparison across conditions were the bias and empirical standard errors of the item difficulties. Bias was defined using the following equation:

$$\text{Bias} = \frac{\sum_r \hat{\delta}_r - \delta}{R}, \tag{2}$$

where $\hat{\delta}_r$ is the estimated item difficulty in replication r, δ is the true generating item difficulty, and R is the number of replications. Bias reflects the average difference of a parameter from the true parameter.

The standard error was defined using the following equation:

$$\text{Standard Error} = \sqrt{\frac{\sum_r (\hat{\delta}_r - \bar{\delta})^2}{R}}, \tag{3}$$

where $\hat{\delta}_r$ is the item difficulty estimate, $\bar{\delta}$ is the mean item difficulty estimate across replications, and R is the number of replications. The empirical standard error reflects the average distance of a parameter from the mean parameter across replications. It is a measure of the variability of a parameter.

Biases and standard errors were calculated for the item difficulties within each condition. Each condition was replicated 1,000 times.

Results

Bias

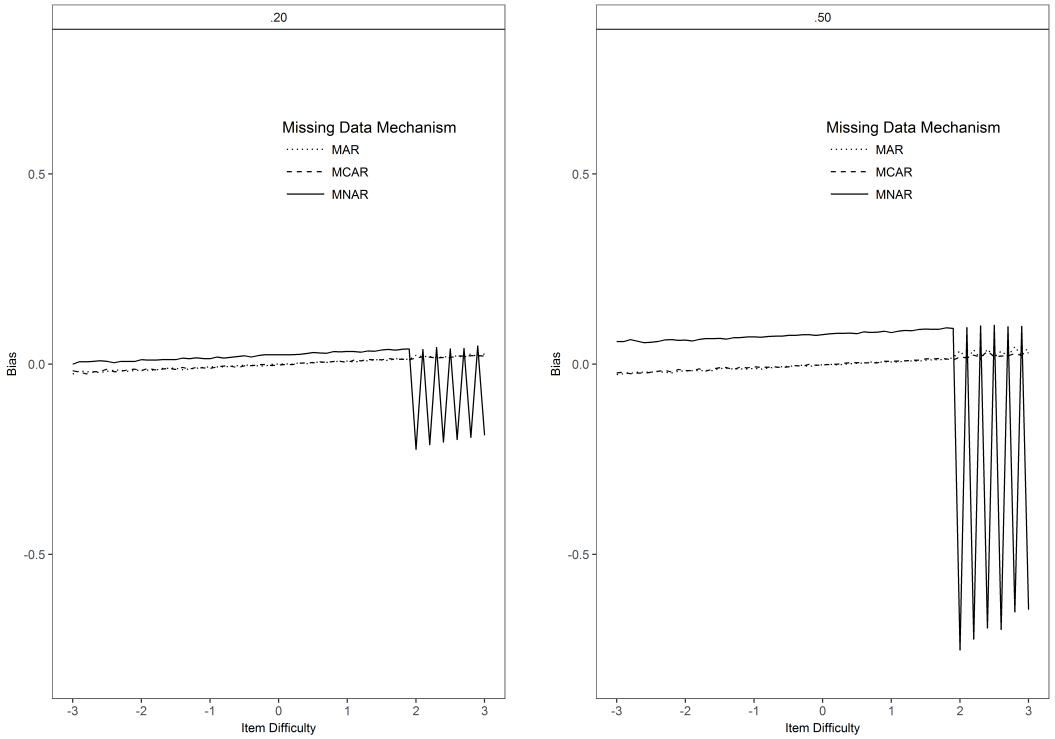
As expected, sample size did not have an effect on item difficulty bias. Therefore, when graphing and discussing bias, results were averaged across sample size. Additionally, the length of the test did not have an effect on bias. Therefore, for simplicity, only the results for the

60 item test will be graphed and discussed. Bias did vary dramatically across the proportion of missing responses factor, and also the missing data mechanism factor. Figures 1a and 1b depict these effects of these factors on the biases of item difficulties.

True item difficulties are on the *x*-axis of Figures 1a and 1b. As discussed in the method, items on the test were arranged by true difficulty, from easiest to hardest. Therefore, the 6 items with missing responses were located at the end of the test. The true difficulties of these items ranged from 2 to 3 logits. Bias of item difficulties comprises the *y*-axis of Figure 1a and 1b. Figure 1a depicts conditions when the proportion of missing responses on each item was .2, and the Figure 1b depicts conditions when the proportion was .5. Each of the three lines represent a different missing data mechanism.

As can be seen from Figures 1a and 1b, item difficulties were negligibly biased when missing responses were MCAR or MAR. This pattern held regardless of whether the proportion of missing responses was .2 or .5. This was expected, as maximum likelihood based estimation should yield unbiased parameters when data are MCAR or MAR. Although not computationally equivalent, proportional curve fitting estimation provides parameters that are functionally equivalent to traditional maximum likelihood, which was confirmed in this study in the case of missing data.

When missing responses were MNAR, item difficulties were severely biased. The bias of items with missing responses was always negative, regardless of the proportion of missing responses or the true difficulty of the item. Thus, items with missing responses were always *underestimated,* meaning they were estimated to be easier than



Figures 1a and 1b. Bias in item difficulties. Results were averaged across sample size. Only results from the 60 item test are displayed in the figure. Each subfigure represents a different proportion of missing data. Figure 1a represents the .20 missingness conditions and Figure 1b represents the .50 missingness conditions. Items, arranged in order of true difficulty, are on the *x*-axis. Biases of the item difficulties are on the *y*-axis. Each line represents a different missing data mechanism.

their true generating item difficulties. The severity of the bias in the MNAR conditions depended on the proportion of missing responses. As can be seen from Figure 1, the underestimation of item difficulties with missing responses was much more dramatic when the proportion of missing responses was .5. When missing responses were MNAR and the proportion of missingness was .5, some items with missing responses were underestimated by more than .6. On the logit scale, this constitutes a substantial degree of underestimation.

Of note is that when data were MNAR and the proportion of missing responses was .5, the 54 items without missing responses were all slightly positively biased. This was due to the method in which the origin of item difficulties was set. To solve the indeterminacy during estimation, it is conventional with Rasch measurement to set the mean of the item difficulties to zero. As stated earlier, items with missing responses were *negatively* biased. Thus, to maintain a mean item difficulty of zero, all of the remaining items with complete data were *positively* biased to compensate. This highlights the nature of parameter calibration using IRT or Rasch measurement. Parameters directly impacted by an undesired circumstance, such as the occurence of missing data or correct guessing, will not only be biased themselves, but also may indirectly bias other parameter estimates in the model. The MNAR results clearly depict this indirect effect. Additionally, recall that only 6 of the 60 items on the test contained missing responses, constituting 10% of the overall items on the test. If this percentage of items had been higher, then the positive bias for the remaining item difficulties was likely to have been even larger than it was in this study.

Standard Errors

As was the case with bias, the standard errors of item difficulties did not change across test length. Therefore, only standard error results for the 60 item test will be discussed. As is the case with almost any statistical model, standard errors were affected by sample size, with larger sample sizes associated with smaller standard errors. The proportion of missing responses and the missing data mechanism both had effects on the standard errors of items with missing responses.

For simplicity, results for the smallest sample size ($N = 500$) and largest sample size ($N = 2{,}000$) will be graphed and discussed. These are displayed in Figures 2a, 2b, and 2c, and Figures 3a, 3b, and 3c respectively.

As with Figures 1a and 1b, true item difficulties comprise the *x*-axis of Figures 2 and 3. Standard errors are on the *y*-axis of both figures. Each a, b, and c subfigure represents a different proportion of missing responses. Each line represents a different missing data mechanism. When missing responses were MCAR or MAR, results were identical. Thus, for graphical simplicity, only results for the MCAR and MNAR mechanisms are displayed. New to Figures 2 and 3 are subfigures that display standard errors when the proportion of missing data was zero, meaning the data was complete. This is represented by the dotted lines in Figures 2a and 3a. The standard errors for the complete data conditions represent the ideal situation and serve as a point of comparison for the various missing data mechanisms.

As can be seen from Figures 2 and 3, the *patterns* of standard errors were equivalent across sample size conditions. Standard errors were smallest for items with a true difficulty near zero. These items were estimated using the most statistical information, given that the examinees were standard normally distributed. As items became easier or more difficult, standard errors increased. Of course, standard errors were always smaller in the large sample size conditions than in the small sample size conditions, as is reflected in the smaller scale of the *y*-axis. However, the *patterns* of standard errors were invariant across sample sizes.

When data were MCAR, items with missing responses had standard errors that were larger than when the data were complete. This effect was more substantial when the proportion of missingness was .5, as opposed to .2. Although not shown, the same was true of the MAR conditions. This was an expected result. When responses are missing, item parameters are estimated using

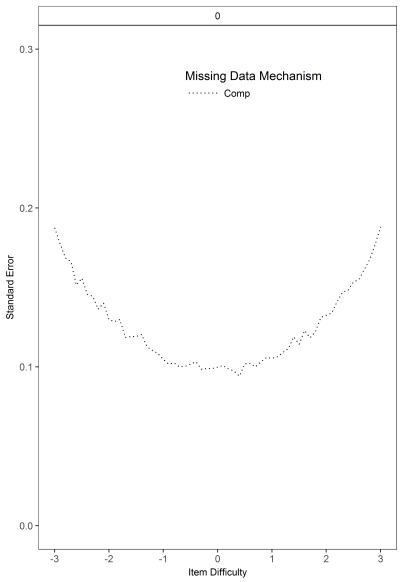


Figure 2a represents conditions when the data was complete, to serve as a point of comparison.

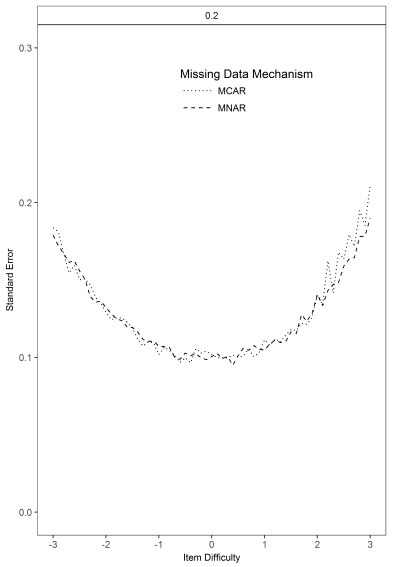


Figure 2b represents the .20 missingness conditions.

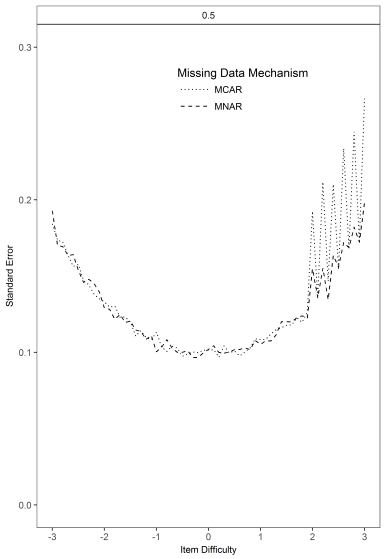


Figure 2c represents the .50 missingess conditions.

Figures 2a, 2b, and 2c. Standard errors of item difficulties when the sample size was 500. Only results from the 60 item test are displayed in the figure. Each subfigure represents a different proportion of missing data. Items, arranged in order of true difficulty, are on the *x*-axis. Standard errors of the item difficulties are on the *y*-axis. Each line represents a different missing data mechanism. Given that MCAR and MAR results were identical, for graphical simplicity, MAR results were omitted from the figure.

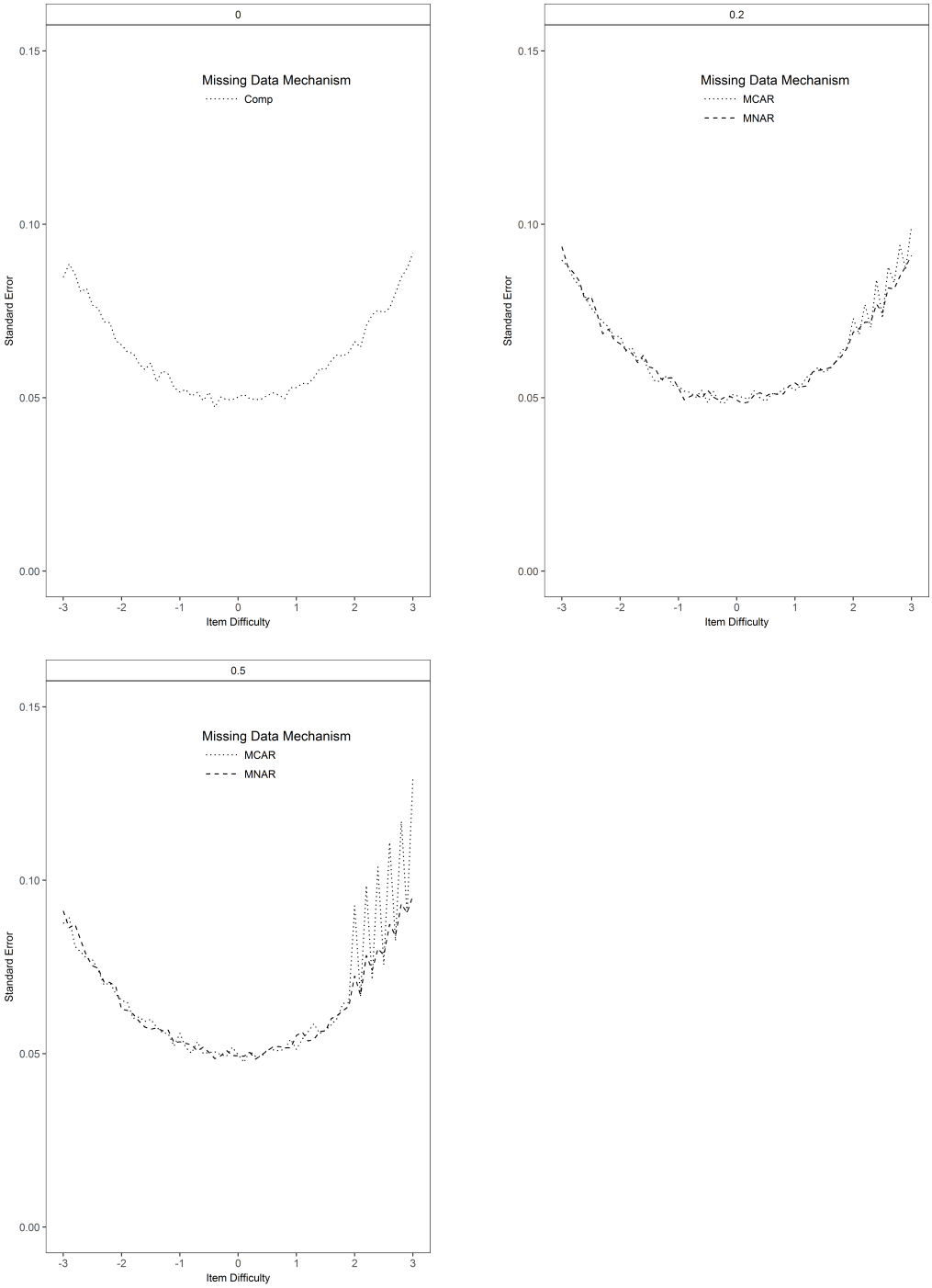


Figure 3a, 3b, and 3c. Standard errors of item difficulties when the sample size was 2,000.

less statistical information than if the data had been complete. As a result, the precision of the parameter estimates decreases and standard errors increase. The larger the proportion of missing data, the less information from which to estimate parameters and thus the larger the standard errors.

An *unexpected* result was the pattern of standard errors when data were MNAR. As can be seen from the solid black line, the standard errors of items from the MNAR conditions were nearly equivalent to the standard errors of the items from the complete data conditions. This pattern held across sample sizes and proportions of missing data. Thus, standard errors for items with missing responses were noticeably smaller when the responses were MNAR as opposed to MCAR or MAR. A reason for this was not immediately apparent. One potential cause of this discrepancy stems from the bias of the item difficulties when responses were MNAR. Recall that items with responses that were MNAR were substantially biased. Perhaps this caused a restriction of range of the difficulties of these items across replications, leading to a decrease in variability and thus a decrease in standard errors. Regardless of the cause, the decrease in standard errors does not make the MNAR mechanism any more appealing, as will be discussed in the following section.

Discussions and Implications

The dichotomous Rasch model is a commonly used measurement model for analyzing responses to multiple-choice test items. In educational and psychological testing, missing data are common. In order for model parameters to be unaffected by missing data, the missingness must be ignorable. Otherwise, parameters may be biased by missing data and thus will be untrustworthy. This simulation study was conducted to explore the ignorability of missing data when using the Rasch model. Specifically, the biases and standard errors of item difficulties were examined under varying conditions of sample size, test length, proportion of missing responses, and the mechanism underlying the missing responses.

Sample size and test length did not have an effect on the biases of item difficulties. The missing data mechanism did have a noticeable effect on item difficulty bias. When missing responses were MCAR or MAR, item difficulties were unbiased. This was true regardless of the proportion of missing responses. When missing responses were MNAR, the items with missingness were negatively biased. The difficulties of these items were underestimated, meaning the items with missing responses were estimated to be easier than implied by their true generating item difficulties. This underestimation of item difficulties when responses were MNAR was moderated by the proportion of missingness. Specifically, bias was much more severe when the proportion of missing responses was .5, as opposed to .2. When the proportion of missing responses was .5, the 54 items with complete data were positively biased in order to compensate for the 6 items with missing responses that were negatively biased.

As was the case with bias, test length did not have an effect on the standard errors of item difficulties. Sample size, as expected, had an effect on standard errors, with larger sample sizes associated with smaller standard errors. However, the pattern of standard errors was equivalent across sample sizes. When missing responses were MCAR or MAR, standard errors for the items with missing responses were larger than if the data had been complete. This effect was magnified as the proportion of missing responses increased. This was expected. When responses are missing, parameters are estimated using less information, which decreases precision and increases standard errors. When responses were MNAR, the standard errors were actually nearly the same as if the data had been complete. This held regardless of the sample size or the proportion of missing responses.

There are several significant findings from this study. The first is that, in terms of bias, the pattern of results were similar to previous studies regarding missing data and maximum likelihood estimation. Recall that Winsteps was used for all parameter estimation in this study. Also recall that Winsteps uses a proportional curve fitting algorithm for estimation, not traditional maximum likelihood using the Newton-Raphson procedure.

Thus, it was encouraging to see that despite the use of a different estimation technique, Winsteps yielded patterns of bias that were consistent with previous literature that employed traditional maximum likelihood. More specifically, it is encouraging that the Rasch model can handle missing responses that are MCAR or MAR, even when half of the responses to an item are missing.

However, a significant finding was the overall detrimental effect that the MNAR mechanism had on *all* item parameters, not just the parameters for items with missing responses. In order to keep the origin of item difficulties at zero, all items with complete data were positively biased to compensate for the negative bias of the items with missing responses. This highlights the tenuous nature of parameter calibration using the Rasch model or any IRT model. A few biased parameters, even when they constitute a small portion of the overall test, can distort the remaining parameters. This is especially problematic with the MNAR mechanism, as there is no way to know, in an applied setting, if the missing responses were indeed MNAR. This is why it is important to take steps to ensure that examinees answer all items. Otherwise, a few "bad apple" items could poison the entire orchard.

Lastly, an unexpected finding was that standard errors in MNAR conditions were smaller than in MCAR or MAR conditions. In fact, they were nearly identical to what the standard errors would have been had the data been complete. In some capacity, this may seem like a benefit of the MNAR mechanism. Indeed, precision of parameter estimates is a desirable quality of any model, not just the Rasch model. However, the dramatic bias of item parameters when missing responses were MNAR outweighed any increase in precision. In essence, the decreased standard errors with MNAR data simply meant that item parameters were more precisely wrong. Therefore, missing responses caused by an MNAR mechanism are undesirable in practice, regardless of increases in precision.

There are several limitations to this study. This was a simulation study, so the "true" item parameters could be known a priori and used as a point of comparison to the estimated parameters. However, there were a limited number of specific conditions from which to gather results. Several potentially interesting factors were not included in this study. For example, regardless of test length, the percentage of items with missing responses was always 10%. A useful follow up to this study would be to explore the effects of varying percentages of items with missing responses on item parameters. It seems likely that as the percentage of items with MNAR missing responses increases, the bias of the remaining items would also increase, in order to maintain the item difficulty origin of zero.

Another limitation is that all item responses in the study were generated to fit the Rasch model, which will not always hold in applied practice. An interesting follow up study would be to determine the effects of various missing data mechanisms on items that did not fit the Rasch model. For example, how would missing data affect estimation of items with a non-zero lower asymptote? Or items with varying levels of discrimination?

A final limitation is that the missing data mechanisms were kept completely separate. Within each condition, all missing responses were generated by a single missing data mechanism. This was done for simplicity and to isolate the effects of each missing data mechanism. However, this would likely be an unrealistic situation in reality. Instead, it is likely that different items on the same test will have different mechanisms underlying missing responses. For example, one item may have missing responses that are MAR, while another may have responses that are MNAR. Additionally, a single item may contain missing responses that constitute multiple missing data mechanisms. For example, one student may inadvertently skip an item, while another student may intentionally skip the same item because it appears too difficult. A useful study would be to examine the effects of multiple types of missing data mechanisms on a single item or test.

References

Andrich, D., and Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement*, *38*, 432-449.

Andrich, D., Marais, I., and Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*, 417-442.

Andrich, D., Marais, I., and Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement*, *76*, 412-435.

Andrich, D., Sheridan, B. E., and Luo, G (2013). RUMM2030: Rasch unidimensional models for measurement [Computer software]. Perth, WA, Australia: RUMM Laboratory.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumaker (Eds.) *Advanced structural equation modeling: Issues and techniques* (pp. 243-277), Mahwah, NJ: Earlbaum.

Enders, C. K., and Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430-457.

De Ayala, R. J., Plake, B. S., and Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*, 213-234.

Enders, C. K. (2005). Estimation by maximum likelihood. *Encyclopedia of Behavioral Statistics*, 1164-1170.

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

Linacre, J. M. (2004). Estimation methods for Rasch measures. In E. V. Smith and R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 25-47). Maple Grove, MN: JAM Press.

Linacre, J. M. (2015). Winsteps (Version 3.92.0) [Computer software and user's manual]. Beaverton, OR: Winsteps.com.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247-264.

Meyer, J. P., and Hailey, E. (2012). A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement*, *13*, 248-258.

Mislevy, R. J., and Wu, P. K. (1988). Inferring examinee ability when some item responses are missing. *ETS Research Report Series*, *1988*, i-75.

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Waterbury, G. T., and DeMars, C. E., (2019). The effects of probability threshold choice on an adjustment for guessing using the Rasch model. *Journal of Applied Measurement*, *20*, 1-12.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., and Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith and R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.