Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 484-506

Rasch Joint Maximum Likelihood Estimation Algorithms and Missing Data

Adam E. Wyse[1]

Abstract

This article examines two approaches for performing joint maximum likelihood estimation with the Rasch model and how these estimation algorithms may be impacted by the amount and type of missing data. The two estimation algorithms include the Newton-Raphson procedure and a proportional curve fitting algorithm. Using simulated data from two different credentialing programs, we found that the amount and type of missing data can impact the amount of error and variability observed in item and person parameters. However, we found that the proportional curve fitting and Newton-Raphson algorithms tended to give virtually identical results. The only differences between the two algorithms were when missing data were created using a computerized adaptive testing algorithm and there were less than 50 scored item responses. In some of these cases, there were very small differences between the two algorithms with the proportional curve fitting algorithm performing slightly better. It is suggested that in most practical applications that one should expect very similar results no matter what algorithm is employed to estimate item and person parameters.

Keywords: Rasch model, estimation algorithms, missing data, joint maximum likelihood

[1] Address Correspondence to: Adam E. Wyse, Ph.D., 1813 Chatham Ave, Arden Hills, MN 55112, Email: adam.wyse[at].renaissance.com

One of the most commonly used models in educational and psychological testing is the Rasch (1960) model. The Rasch model posits that the probability of obtaining a correct response on an item is a function of the person's ability and the difficulty of the item. A key aspect of applying the Rasch model to exam data is obtaining estimates of item and person parameters. There are many different approaches for obtaining estimates of item and person parameters, including joint maximum likelihood estimation (JMLE), conditional maximum likelihood (CMLE), marginal maximum likelihood estimation (MMLE), the PROX method, the PAIR method, and Bayesian estimation methods (see Engelhard, 2014; Linacre, 1999; 2004; Molenaar, 1995). Among these methods, JMLE is one of the most commonly used methods due in large part to its widespread availability in several software packages, such as Winsteps (Linacre, 2016), Facets (Linacre, 2014), ConQuest (Adams, Wu, & Wilson, 2012), jMetrik (Meyer, 2016), mixRasch (Willse, 2014), and TAM (Kiefer, Robitzsch, & Wu, 2016).

Although JMLE is used in many different software programs, the way that JMLE is carried out is not always the same across programs. Some software packages use proportional curve fitting algorithms, while other software packages use Newton-Raphson algorithms. In discussing the choice to implement a proportional curve fitting algorithm in Winsteps, Linacre (2016) explains that Newton-Raphson estimation has proven to be unstable with sparse data sets and odd score distributions. This statement implies that knowing the algorithm implemented by the software and the properties of the data one is employing it on may be important.

Despite the statement in the Winsteps manual on the negative performance of the Newton-Raphson algorithm and the robust performance of the proportional curve fitting algorithm with sparse data, published research to substantiate this claim is lacking in the literature. The authors performed a literature search and did not find a single published article that examined how the type and amount of missing data may impact different JMLE algorithms. This article examines how the amount and type of missing data may impact these JMLE algorithms using two real data-based simulations. Our specific research questions are:

1) How does the type and amount of missing data impact different JMLE algorithms?
2) Are results consistent across different datasets, or does utilizing different data produce disparate results?

In the next section, we outline the two different JMLE algorithms that are the focus of our investigations. Then, we review prior research on missing data and how it may impact the estimation of item and person parameters for the Rasch model. The next section provides a description of the data and methods used to investigate our research questions. The results for the different JMLE algorithms are then compared for two different datasets. The article concludes with discussion of results and suggestions on using JMLE when there may be missing data.

Joint Maximum Likelihood Estimation Algorithms

JMLE was first introduced to estimate Rasch item and person parameters by Wright and Panchapsken (1969). JMLE derives its name from the fact that item and person parameters are jointly estimated. In the context of the Rasch model, this means that estimates for person parameters are used to estimate item parameters and then the estimates for items are used to estimate person parameters. This process is iterated until the convergence level is reached. It is well known that JMLE can give biased estimates. Generally, parameters improve when the numbers of items and people increase, and the sample used to estimate the model well represents the population of interest (Linacre, 1999, 2004; Meyer & Hailey, 2012; Svetina et al., 2013; Wang & Chen, 2005; Wright & Douglas, 1977; Wright, 1988; Wyse & Babcock, 2016). JMLE estimates are also known to be inconsistent (Anderson, 1973; Del Pino, San Martin, González, & De Boeck, 2008; Ghosh, 1995; Haberman, 1977; Jansen, van den Wollenberg, & Wierda, 1988; Linacre, 1999, 2004). Despite these challenges, JMLE remains a staple in estimating Rasch item and person parameters and yields acceptable results in most practical applications.

There are several different options to compute the item and person parameters using JMLE. All the algorithms have the same starting point. In particular, each algorithm is focused on estimating item and person parameters for the Rasch model, which can be written as:

$$P_{ig} = P\left(X_i = 1 \middle| \beta_g, \delta_i\right) = \frac{e^{\beta_g - \delta_i}}{1 + e^{\beta_g - \delta_i}} \tag{1}$$

where β_g is the person measure of person g and δ_i is the difficulty parameter for item i (Rasch, 1960). All the algorithms also need initial estimates of the item difficulty and person parameters to begin the iteration process. An initial estimate of the item parameter for an item is typically found as:

$$\hat{\delta}_i^{(0)} = \log\left(\frac{N_i - s_i}{s_i}\right) - \frac{\sum_{i=1}^{n} \log\left[\left(N_i - s_i\right)/s_i\right]}{n}\ , \tag{2}$$

where N_i is the total number of people responding to item i, n is the total number of items, and s_i is the number of people correctly answering item i (Wright & Panchapakesan, 1969). Similarly, an initial estimate of the person parameter is typically found as:

$$\hat{\beta}_g^{(0)} = \log\left(\frac{r_g}{n_g - r_g}\right), \tag{3}$$

where r_g is the total number of items answered correctly by person g and n_g is the number of items that they answered (Wright & Panchapakesan, 1969). It is important to recognize that the total number of persons responding to an item and the total number of items answered by a person may differ across items and people.

The algorithms differ in how the item and person parameters are found during the iteration process. In the Newton-Raphson procedure, the derivatives of Equation 1 with respect to items or persons are used to determine the new item and person parameters. New item parameter estimates are obtained as:

$$\hat{\delta}_i^{(t+1)} = \hat{\delta}^{(t)} - \left(\frac{s_i - \sum_{g=1}^{N} \hat{P}_{ig}}{\sum_{g=1}^{N} \hat{P}_{ig}\left(1 - \hat{P}_{ig}\right)}\right)^{(t)}, \tag{4}$$

where t is the iteration, N is the total number of people answering the item, s_i is the number of people correctly answering item i, and $\hat{P}_{ig}$ is the estimated probability of correct response to the item based on the current person and item parameter estimates from Equation 1 (Wright & Panchapakesan, 1969). These estimates are then centered at zero and used to find new person parameter estimates using the equation:

$$\hat{\beta}_g^{(t+1)} = \hat{\beta}_g^{(t)} - \left(\frac{r_g - \sum_{i=1}^{n} \hat{P}_{ig}}{\sum_{i=1}^{n} \hat{P}_{ig}\left(1 - \hat{P}_{ig}\right)}\right)^{(t)}, \tag{5}$$

where the terms have the same meaning as described above (Wright & Panchapakesan, 1969). The new estimates are then compared with the estimates from the previous iteration for both items and people and if the differences between the new and old estimates are smaller than the desired convergence level the iterative process stops. Otherwise, another iteration of the process is performed. The Newton-Raphson procedure is efficient and works quite well in most applications. However, it is possible

for the Newton-Raphson procedure to run into convergence problems in some situations (Linacre, 1987; 2004; Molenaar, 1995).

In the proportional curve fitting procedure, one tries to approximate the item and test characteristic curves for the Rasch model using linear equations. The item characteristic curve is Equation 1 and the test characteristic is the sum of item characteristic curves over items. To implement this algorithm, one must first define a starting deviation measure, $d^{(0)}$. This deviation measure defines the length of the line segments that are fit to the characteristic curves. Typically, $d^{(0)}$ is set to 1 as a starting value. To figure out the item parameter in iteration *t*, one computes

$$\hat{\delta}_i^{(t+1)} = \delta^{(t)} + \frac{-d^{(t)}}{\left(\log\left(\frac{N_i - \sum_{g=1}^{N}\hat{P}_{ig}}{\sum_{g=1}^{N}\hat{P}_{ig}}\right) - \log\left(\frac{N_i - \sum_{g=1}^{N}\hat{P}^{*}_{ig}}{\sum_{g=1}^{N}\hat{P}^{*}_{ig}}\right)\right)}\log\left(\frac{N_i - s_i}{s_i}\right) + \frac{d^{(t)}\log\left(\frac{N_i - \sum_{g=1}^{N}\hat{P}_{ig}}{\sum_{g=1}^{N}\hat{P}_{ig}}\right)}{\left(\log\left(\frac{N_i - \sum_{g=1}^{N}\hat{P}_{ig}}{\sum_{g=1}^{N}\hat{P}_{ig}}\right) - \log\left(\frac{N_i - \sum_{g=1}^{N}\hat{P}^{*}_{ig}}{\sum_{g=1}^{N}\hat{P}^{*}_{ig}}\right)\right)} \quad (6)$$

where $\hat{P}^{*}_{ig}$ is the value of Equation 1 with $\hat{\delta}^{(t)} + d^{(t)}$ input for the item parameter and the other terms have the same meaning as before (Linacre, 2016). These item difficulty parameters are then centered at zero and used to find new person parameters using the equation:

$$\beta_g^{(t+1)} = \hat{\beta}_g^{(t)} + \frac{-d^{(t)}}{\left(\log\left(\frac{\sum_{i=1}^{n}\hat{P}_{ig}}{n_g - \sum_{i=1}^{n}\hat{P}_{ig}}\right) - \log\left(\frac{\sum_{i=1}^{n}\hat{P}^{\#}_{ig}}{n_g - \sum_{i=1}^{n}\hat{P}^{\#}_{ig}}\right)\right)}\log\left(\frac{r_g}{n_g - r_g}\right) + \frac{d^{(t)}\log\left(\frac{\sum_{i=1}^{n}\hat{P}_{ig}}{n_g - \sum_{i=1}^{n}\hat{P}_{ig}}\right)}{\left(\log\left(\frac{\sum_{i=1}^{n}\hat{P}_{ig}}{n_g - \sum_{i=1}^{n}\hat{P}_{ig}}\right) - \log\left(\frac{\sum_{i=1}^{n}\hat{P}^{\#}_{ig}}{n_g - \sum_{i=1}^{n}\hat{P}^{\#}_{ig}}\right)\right)} \quad (7)$$

where $\hat{P}^{\#}_{ig}$ is the value of Equation 1 with $\hat{\beta}^{(t)} + d^{(t)}$ input for the person parameter and the other terms have the same meaning as before (Linacre, 2016). The new estimates are then compared with the estimates from the previous iteration for both items and people and if the differences are smaller than the convergence level the iteration process stops. Otherwise, another iteration is performed with the maximum difference between any two parameters becoming the new deviation measure. Like the Newton-Raphson procedure, the proportional curve fitting algorithm is efficient and works well in most applications. Based on Linacre (2016), one would infer that this approach should be more robust than the Newton-Raphson procedure when there is missing data.

Missing Data and the Rasch Model

There are three kinds of missing data that are encountered in practice; missing completely at random (MCAR), missing at random (MAR), and missing not at random (NMAR) (Dempster, Laird, & Rubin, 1977; Little & Rubin, 2002). MCAR data occur if the missing data do not depend on the values of the observed and unobserved data. MAR data occur if the missing data only depend on the values of the observed data and do not depend on the values of the unobserved data. NMAR data occur if the missing data depend on the values of the unobserved data. Research on the Rasch model and how it is impacted by missing data has looked at all three kinds of missing data. This research has followed three different lines of inquiry.

The first line of inquiry has focused on developing or extending existing Rasch models to handle various kinds of missing data. Examples of this type of research include Verhelst and Glas (1993), Holman and Glas (2005), Rose, von Davier, and Nagengast (2010), and Bertoli-Barsotti and Punzo (2013). This line of research often does not look at how different estimation algorithms are impacted by missing data, but it does suggest that the kind of missing data can impact results and that using a model that considers the kind of missing data can improve parameter estimates. Given that the kind of missing data can impact results, we look at how the estimation algorithms may be impacted by four types of missing data in our simulations.

The second line of inquiry looks at how different ways of handling missing data may impact parameter estimates. For example, Hohensinn and Kubinger (2011) and Custer, Sharairi, and Swift (2012) looked at how treating data as missing versus incorrect may impact Rasch model results. Both studies showed that scoring omitted and not reached items as incorrect can lead to biased results. Shin (2009) showed that treating responses as missing was preferable to treating responses as incorrect when performing Rasch-based true-score equating. Ludlow and O'Leary (1999) also found that different methods of scoring omitted and not reached items can lead to different results for a cognitive ability test. Sijtma and van der Ark (2003) looked at how various imputation techniques may impact results and found that a strategy based on response function imputation tended to work quite well. This line of inquiry seems to suggest that how one handles missing data can impact the parameters that are obtained. Since our concern is with how the two estimation algorithms are impacted by missing data, we examine the case where missing data are treated as missing. We do not examine how results may be impacted by scoring items as incorrect, using listwise deletion, or employing different imputation strategies.

The third line of inquiry examines how different estimation strategies may be impacted by missing data. Heine and Tarnai (2015) developed a pairwise estimation algorithm and looked at how the algorithm compared to MMLE, CMLE, and an imputation-based strategy for an eight-item survey with various levels of missing data. Their research showed that the pairwise estimation algorithm worked quite well, especially as the amount of missing data was increased. However, the authors only used a single replication, focused on one data set with a small number of items, and did not

include JMLE as part of their investigations. Andrich and Luo (2003) also developed a conditional pairwise estimation algorithm that uses principal components and can handle missing data and low frequency counts. They showed that this method performed well in a simulation study. DeMars (2002) used simulated data to compare JMLE and MMLE and showed that the two methods can produce different results when data are not MAR. In this article, we examine how JMLE algorithms work with two real data sets where the ways that data are set to missing and the amount of data set to missing are varied in different conditions.

Data and Methods

Data for this study come from two medical imaging credentialing programs. The exams are continuously administered via computer throughout the year. Passing the exams and earning a credential is often a requirement to obtain employment in each discipline. Examinees have up to three attempts to pass the exams. Each exam program uses the Rasch model for scoring and equating purposes, and estimation of item and person parameters is carried out using JMLE with Winsteps (Linacre, 2016). For the purposes of scoring and equating the exams, all calibrations are based on responses from first-time candidates. Examinees who repeat the exam are not included in the calibration sample.

Table 1 shows the number of scored items, the number of first-time candidates, the average Rasch item difficulty parameters, the standard deviation of the item difficulty parameters, the average person measures, and the standard deviation of the person measures for one form for the two exam programs. The Rasch item difficulty and person estimates in Table 1 were calculated using a proportional curve fitting algorithm, but the results were identical using the Newton-Raphson algorithm. One can see that Program 1 had a much lower volume of first-time candidates taking the form. Each examinee had scored responses on all 160 items for the first program and all 165 items for the second program. The average Rasch item difficulty parameters for each form were equal to 0.00 since this is the constraint used to identify the initial scale when calibrating data. The standard deviation of the item difficulty parameters was similarly around 1.00 for both programs. The average Rasch person measures were similar between the two programs with Program 1 having slightly lower average person measures. As is common with credentialing exams, the average person measures were notably higher than the average item difficulty of the forms, indicating that many candidates did well on the items. We chose to use these two data sets in our simulations because they represent ranges of sample sizes with which the Rasch model is used, and they illustrate common situations in which the Rasch model is applied to exam data in practice.

Table 1:

Summary Statistics for Two Credentialing Exam Programs

Exam Program	Number of Scored Items on Form	Number of Candi-dates Taking Form	Average Rasch Item Diffi-culty	SD of Rasch Item Diffi-culty	Average Rasch Person Measure	SD of Rasch Person Measure
Program 1	160	300	0.00	1.02	1.29	0.68
Program 2	165	1817	0.00	1.00	1.36	0.62

The first factor that we considered in our simulations was the type of missing data. The first kind of missing data included randomly selecting a number of item responses to be treated as scored and setting the rest of the responses to missing. This condition simulates a set of missing data that are designed to be missing completely at random. The second kind of missing data was created by keeping a certain number of scored item responses at the beginning of the exam and setting the rest of the item responses at the end of the exam to missing. This condition simulates data that might be impacted by item order effects, where items at the end of the test are less likely to be answered than items at the beginning of the test. There might not be a lot of item order effects for these data since the order of the items for each candidate is randomized as a test security measure. The randomization means that candidates see the same set of scored items on a form, but the order in which the items appear generally differs across candidates. If the item order effects are not large, the results for the first and second condition will be similar. The third condition applied a computerized adaptive testing (CAT) algorithm to determine the scored and missing data for each examinee. In this case, we used the catIrt (Nydick, 2014) package and we simulated fixed length adaptive tests for each examinee where five items were randomly given at the start of the exam, and then items were selected based on maximizing the information at the person's current ability estimate. We selected the items from 160 or 165 scored item responses observed for that person. This condition simulates a common type of missing data in large-scale assessments, where item responses may be missing based on the person's ability, item difficulty, and scored responses to previous items. The fourth condition simulated data by setting item responses to missing for people who were in the bottom quartile of the score distribution. In this case, we randomly selected the scored responses and we set the rest of the responses to missing. This condition simulates a condition where the missing data is fully dependent on the ability of the person with low ability candidates providing fewer item responses. It is important to look at the type of missing data because previous research indicates that the kind of missing data may impact Rasch item and person parameter estimates.

The second factor that we considered was the number of scored item responses. The number of scored item responses included 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 scored items responses. The number of scored item responses is inversely related to

the number of missing responses, which means that when the number of scored responses is less the number of missing responses is greater and vice versa. It is important to look at the number of scored item responses because the number of scored item responses may impact the performance of different algorithms. In particular, one may expect more error in item and person parameters when the number of scored item responses are fewer.

These two factors were fully crossed to produce 40 simulated conditions for each exam program. For each simulated condition, expect for the conditions that involved the item order effects, we ran 100 replications. For the conditions that involved the item order effects, we ran a single replication. We only ran a single replication in these conditions because the order of the items did not change, which means that if we ran multiple replications we would have ended up with the same data. For each simulated data set, we estimated the item and person parameters using the Newton-Raphson and the proportional curve fitting algorithm. Both algorithms were written in R (R core team, 2014) and used a convergence criterion of 0.01. For people that obtained the maximum possible score for a set of scored items we set their ability estimates to 6 and for people that obtained the minimum possible score we set their ability estimates to -6. Similarly, for items that everyone got correct we set the item difficulty estimates to 6 and for items that everyone got incorrect we set the item difficulty estimates to -6.

To evaluate the performance of the two different estimation algorithms, we looked at three different indices for each estimation algorithm and each condition. The first index we computed was the mean absolute difference between the estimated item or person parameters for the simulated data set and the values obtained with the full set of data with no missing data. For items, this statistic can be represented as:

$$MAD = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}|\hat{\delta}_{ir}-\hat{\delta}_{i}|}{R\times n}, \tag{8}$$

where $\hat{\delta}_{ir}$ is the item difficulty estimates for item *i* for replication *r*, $\hat{\delta}_{i}$ is the item difficulty estimate for item *i* for the full sample of population, *n* is the number of items, and *R* is the number of replications. For people, this statistic can be represented as:

$$MAD = \frac{\sum_{r=1}^{R}\sum_{g=1}^{N}|\hat{\beta}_{gr}-\hat{\beta}_{g}|}{R\times N}, \tag{9}$$

where $\hat{\beta}_{gr}$ is the person parameter estimate for person *g* for replication *r*, $\hat{\beta}_{g}$ is the person parameter estimate for person g for the full sample of items, *N* is the number of people, and *R* is the number of replications. The goal is that the MAD is close to zero because this indicates that the item or person parameter estimates for that replication are close in absolute value to the item or person parameter estimates in the full

data set. We rounded the MAD statistic to two decimal places when computing it for each algorithm and condition.

The second index we computed was the root mean squared difference between the estimated item or person parameters for the simulated data set and the estimated values from full set of data with no missing data. For items, this statistic can be represented as:

$$RMSD = \sqrt{\frac{\sum_{r=1}^{R}\sum_{i=1}^{n}\left(\widehat{\delta}_{ir}-\widehat{\delta}_{i}\right)^{2}}{R\times n}}, \tag{10}$$

where the terms have the same meaning as before. For people, this statistic can be represented as:

$$RMSD = \sqrt{\frac{\sum_{r=1}^{R}\sum_{g=1}^{N}\left(\widehat{\beta}_{gr}-\widehat{\beta}_{g}\right)^{2}}{R\times N}}, \tag{11}$$

where the terms have the same meaning as before. The goal is that the RMSD is close to zero because this indicates that the squared differences between the item or person parameter estimates for that replication and those in the full data set were small. Similar to the MAD statistic, we rounded the RMSD statistic to two decimal places when computing it. We focused on the MAD and RMSD as opposed to the mean difference because the mean of the item parameters is set equal to zero to identify the scales. This means that the mean difference for items would be uniformly equal to zero across conditions.

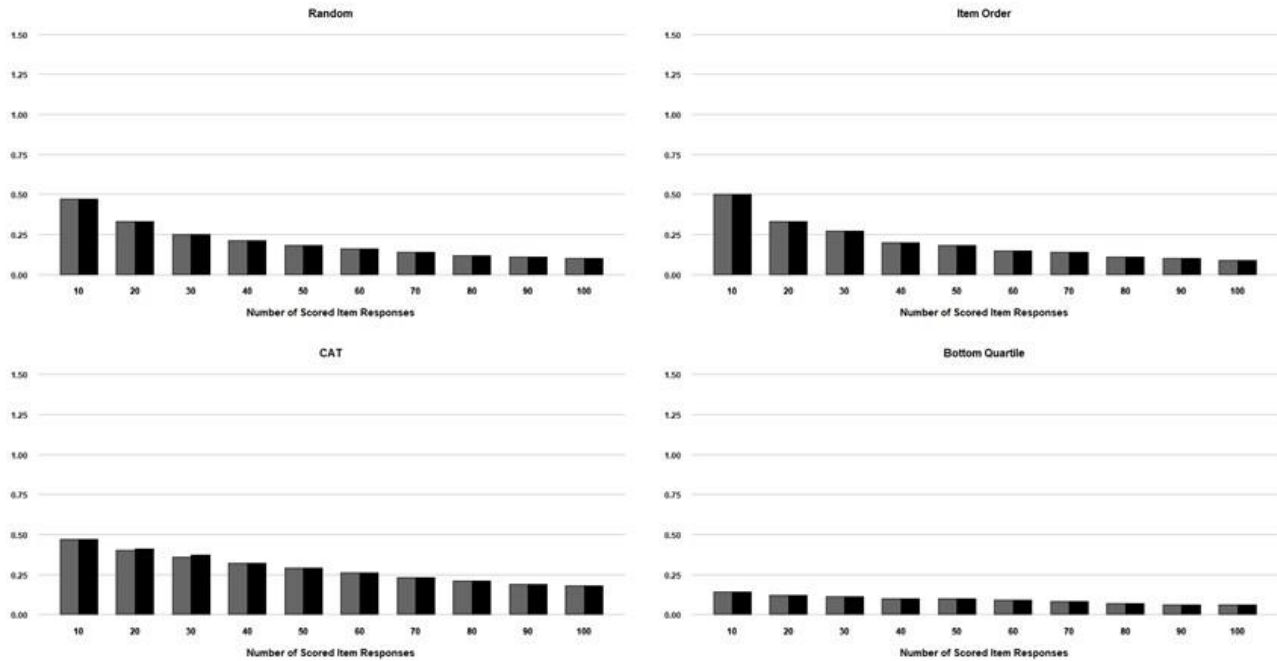
The third index we looked at was the average standard deviation of the estimated item or person parameters over replications. This statistic helps provide an indication of the variability of the item or person parameters. The goal is that the average standard deviation of the estimated item or person parameters over replications would be close to the standard deviations reported in Table 1 for each program.

Based on what is stated in the Winsteps manual (Linacre, 2016), one would anticipate that the MAD, RMSD, and standard deviations would differ somewhat for the proportional curve fitting and Newton-Raphson algorithms, and the proportional curve fitting algorithm would outperform the Newton-Raphson algorithm especially when the number of scored item response are fewer (i.e., the amount of missing data is higher). One would also expect to find differences based on the type of missing data as previous research indicates some differences based on the type of missing data investigated. In addition, one would anticipate higher values of MAD and RMSD and more variability when the number of scored item responses is less.

Results

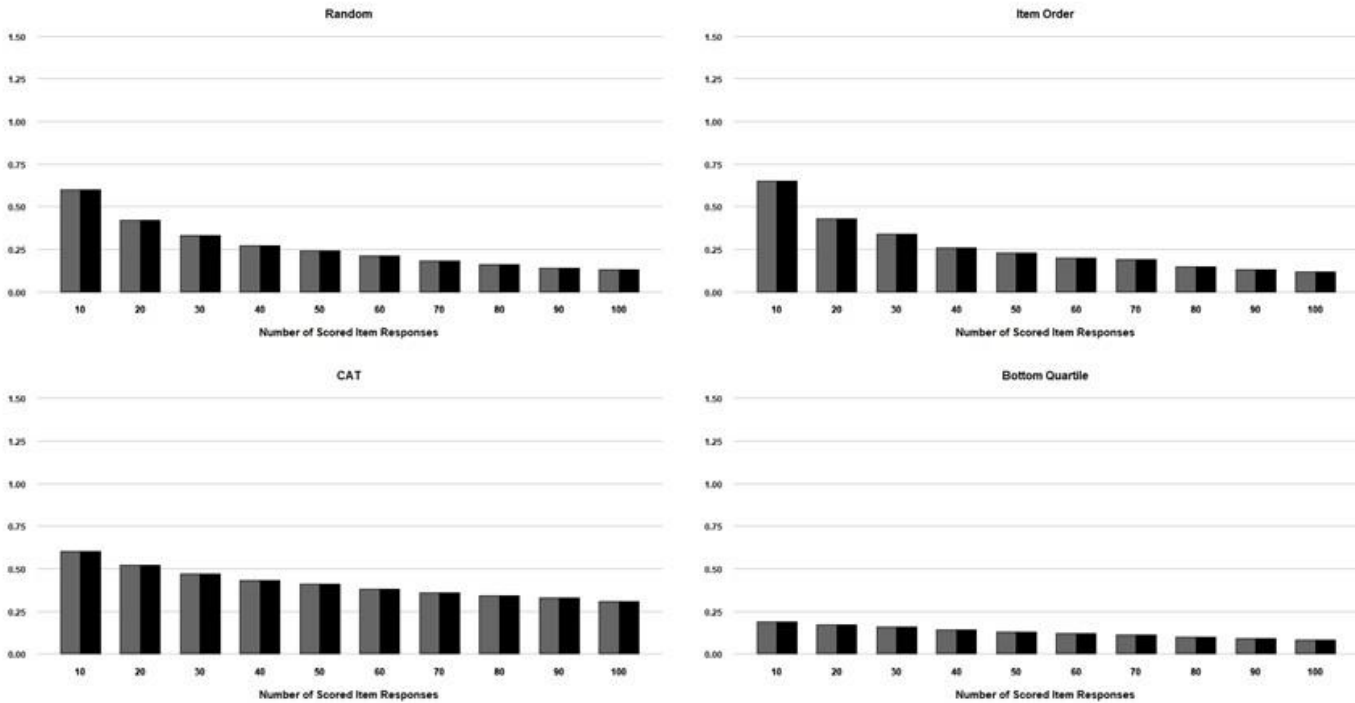
Figures 1, 2, and 3 show the MAD, RMSD, and average standard deviations for the four types of missing data for the item parameter estimates for the first exam program. One can see several trends in the results. First, one can see that the number of scored responses appeared to impact the MAD, RMSD, and average standard deviations with the statistics being higher when there were fewer scored item responses. Second, one can see some differences in the values of the statistics based on the type of missing data. In particular, one can see that when the data were set to missing for people in the bottom quartile of the score distribution that the values of the MAD and RMSD statistics were lower and the average SD was closest to the value of 1.02. Randomly setting responses to missing tended to result in the next lowest values. Using a CAT algorithm generally had the highest values, except when there were only 10 items responses in which case setting responses at the end of the test to missing had the highest values. One might find it a little surprising that randomly setting item responses to missing did not perform the best across all conditions. However, the fact that the method based on setting item responses in the bottom quartile produced the best results makes some sense because this condition had about a ¼ the amount of missing data compared to other conditions. The lower amount of missing data appears to compensate for the fact that missing data came from the bottom quartile of the score distribution. Another key finding is that values for the proportional curve fitting algorithm and Newton-Raphson algorithm were very similar for a given number of scored item responses and specific type of missing data across all three figures. In fact, there were only a few cases when employing CAT algorithms where there were any differences. These differences occurred when there were 20 or 30 scored item responses for the MAD statistic with the proportional curve fitting algorithm performing slightly better than the Newton-Raphson algorithm.

Figure 1: MAD for Item Parameter Estimates for Simulation Using First Credentialing Program



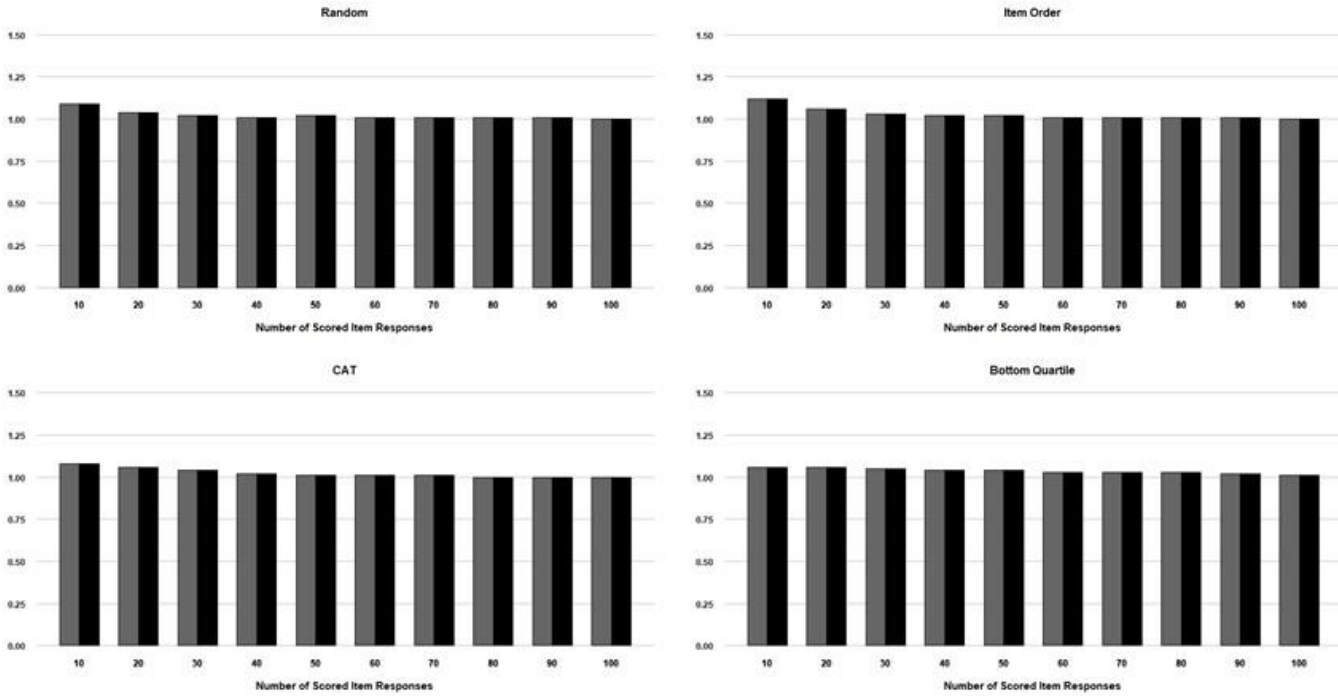
Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

Figure 2: RMSD for Item Parameter Estimates for Simulation Using First Credentialing Program



Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

Figure 3: SD for Person Parameter Estimates for Simulation Using First Credentialing Program

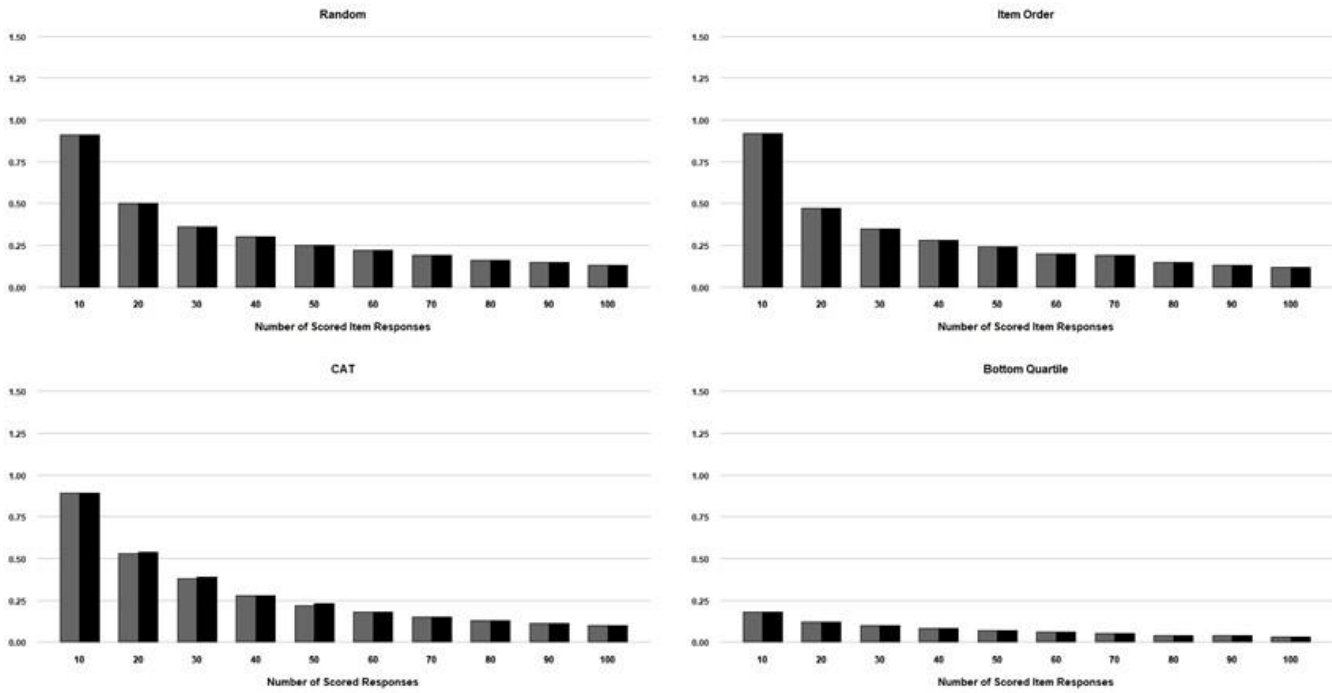


Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

Figures 4, 5, and 6 show the results for the four types of missing data for the person parameter estimates for the MAD, RMSD, and average standard deviations, respectively, for the first exam program. Some of the patterns observed with the item parameter estimates also held for the person parameter estimates. In particular, results suggested that number of scored item responses impacted the MAD, RMSD, and average standard deviation statistics with fewer scored item responses yielding larger values of the statistics. We also found that the type of missing data appeared to impact the values of the statistics with the bottom quartile data having lower RMSD and MAD statistics and average standard deviations that best matched the value of 0.68 shown in Table 1. In addition, we again found that the only type of data where there were any differences between the proportional curve fitting and Newton-Raphson algorithms was when the missing data were created using a CAT algorithm. These differences occurred for MAD and RMSD statistics in a few cases where there were less than 50 scored items responses, but again the differences between the two algorithms were very small.

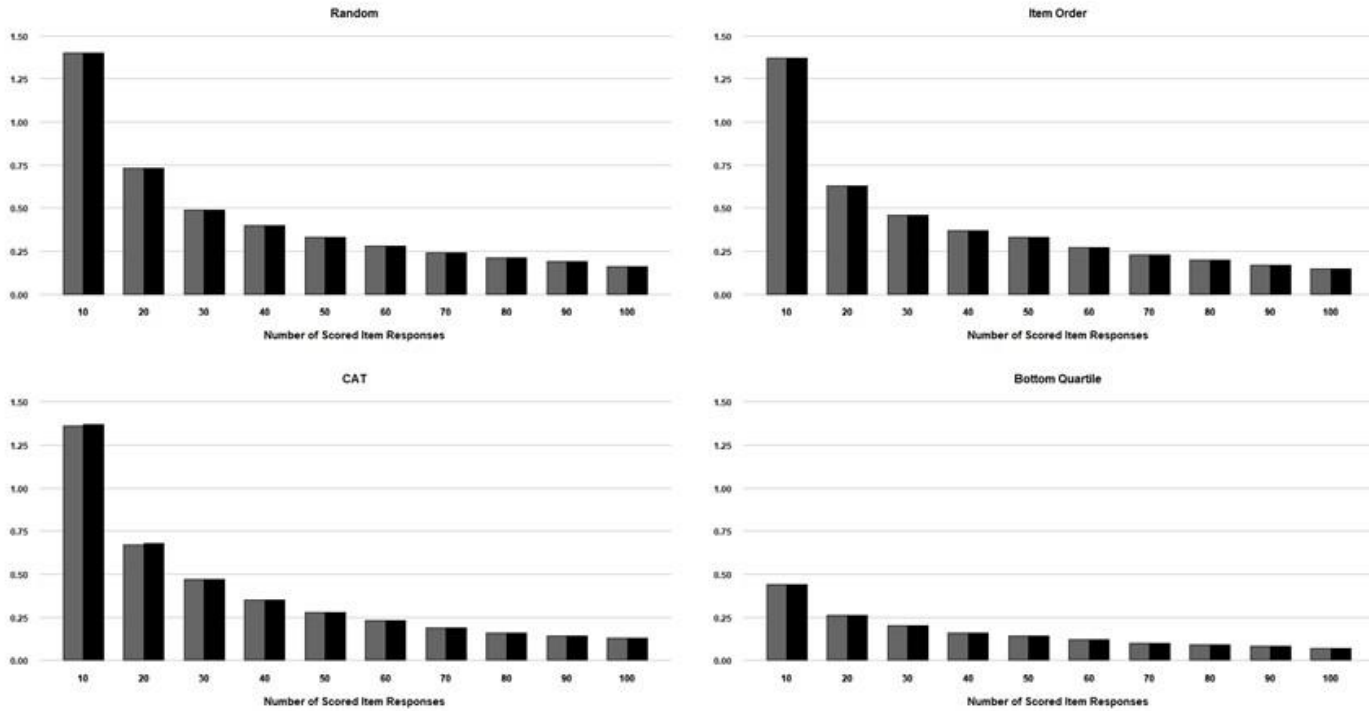
There were two key notable differences in the results for the person parameters compared to the item parameters. First, the person parameter estimates were less accurately estimated than the item parameter estimates. Second, we found that creating missing data by random setting item responses to missing in several cases led to the largest values of the MAD and RMSD, especially with higher numbers of scored responses. These results may seem odd of the surface. However, they do make some sense as one thinks more about how data are created for the other three methods. The bottom quartile again has the least amount of missing data, which help explains why this method performed better. The CAT algorithm works differently than randomly setting responses to missing and specifically selects what items to keep based on the ability level of each examinee. The targeting of items generally results in greater precision of person parameters for CATs, which helps explain why the CAT algorithm works better. The item order method appears to perform slightly better than the random method because as the test gets closer to the end examinees are more likely to randomly guess and spend less time on the items than at the beginning of the test. Removing responses at the end of the test, as happens with the item order method, appears to slightly improve estimation of person parameters because some of these odd responses are more likely to be removed when estimating person parameters. This can make person parameters a little more precise for the item order method than the random method for the same number of scored item responses.

Figure 4: MAD for Person Parameter Estimates for Simulation Using First Credentialing Program



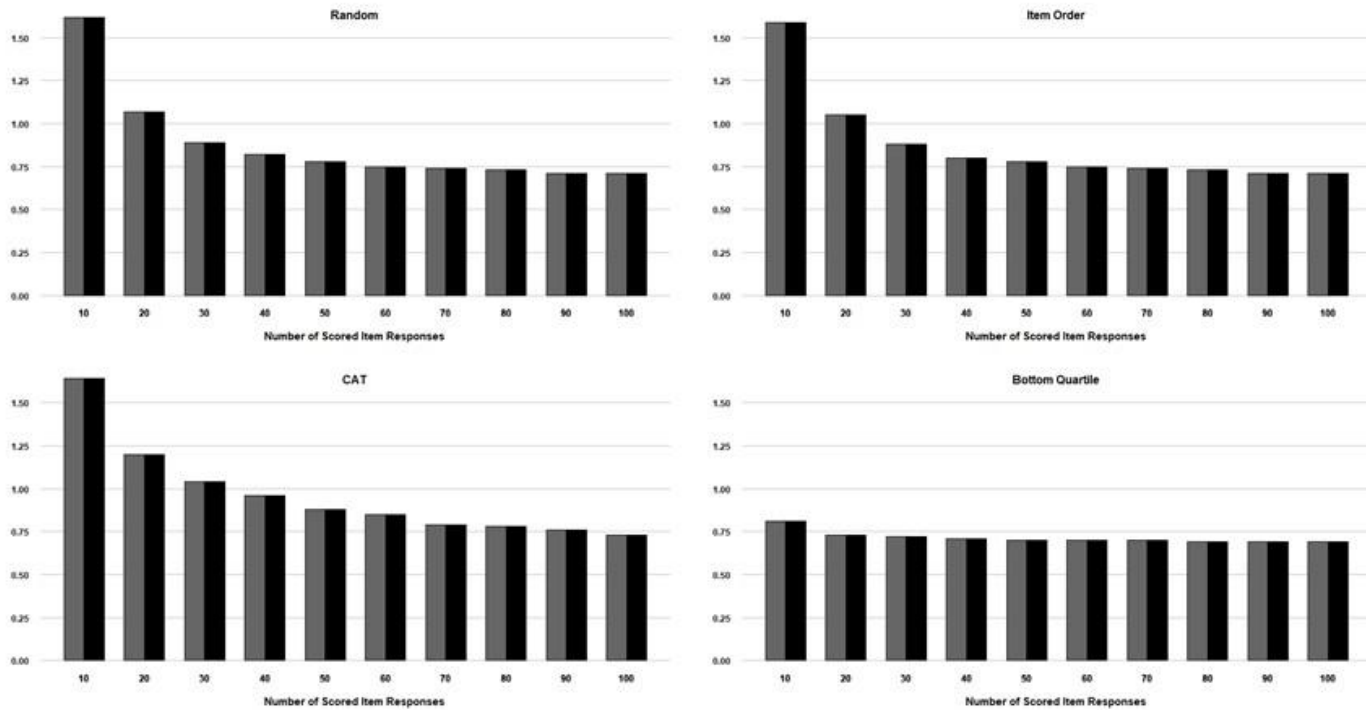
Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

Figure 5: RMSD for Person Parameter Estimates for Simulation Using First Credentialing Program



Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

Figure 6: SD for Person Parameter Estimates for Simulation Using First Credentialing Program



Note. The gray bars show results for the proportional curve fitting method and the black bars show results for the Newton-Raphson method.

The trends for the first credentialing program played out with the second credentialing program as well. In particular, we found that the number of scored item responses impacted the MAD, RMSD, and average standard deviations statistics for both item and person parameter estimates with fewer scored item responses leading to higher values of the statistics. We also found that the type of missing data had some impact with the bottom quartile data producing the lowest values of the statistics. Like with the first program, we also found some very small differences between the proportional curve fitting algorithm and the Newton-Raphson algorithm when using CAT algorithms to create the missing data with the differences occurring when there were less than 50 scored item responses. The proportional curve fitting algorithm again performed a little bit better. The biggest differences in results for the second program were that the values of the MAD and RMSD for the item parameter estimates were notably less because more examinees took the exam. For example, for 10 scored item responses the values MAD statistics were 0.24 for the random data, 0.22 for the item order effect data, 0.24 for the CAT data, and 0.10 for the bottom quartile data. The average standard deviations were also closer to the values in Table 1 for the second program than for the first program. Because of space considerations and because the trends in results were the same with the second program as they were with the first program, we decided not to include figures showing the results for the second program in the main text of the article.

Discussion and Conclusion

The purpose of this article was to explore how the amount and type of missing data impacted two different JMLE algorithms. Using simulated data from two credentialing programs, we found that the amount of missing data impacted the error present in item and person parameter estimates with more error found when there were fewer scored item responses. The average standard deviations also tended to be greater when there were fewer scored item responses. We also found that the type of missing data could impact the error present in item and person parameters with less error found when data were created by only setting responses to missing for people in the bottom quartile of the score distribution. We found that randomly setting item responses to missing tended to work better than setting item responses to missing using a CAT algorithm or setting item responses to missing at the end of the test for item parameter estimation. The same finding did not hold for person parameter estimation, where using a CAT algorithm and deleting item responses at the end of the test often performed slightly better than randomly setting item responses to missing. The fact that randomly setting item responses to missing worked better for item parameter estimation and worse for person parameter estimation than using a CAT algorithm or setting item responses to missing at the end of the test can be explained by the fact that randomly setting item responses to missing tends to give a better match to the item responses of the full population, which is generally important when figuring out item parameter estimates. However, for person parameter estimation, deleting item responses that are less targeted to the person's ability level or that may have been

guessed or given with less effort can lead to better ability estimates when the number of item responses are fewer. We also found, contrary to what is stated in Linacre (2016), that proportional curve fitting and Newton-Raphson algorithms performed similarly in most cases. In fact, the only differences we found where when CAT algorithms were used to create the missing data and the number of scored item responses were less than 50. In some of these cases, the proportional curve fitting algorithm performed slightly better, although the differences between algorithms were not practically significant.

Of course, an important question is whether our simulated conditions were representative of results that may be expected in other circumstances. It is possible that our simulations may not have captured all types of missing data or situations that may be encountered in practice. For example, it is possible that the shape of the score distribution in other situations may be different than the ones included in our analyses. It is also possible that the number of items or people may differ. To evaluate some of these possibilities, we ran other simulations with several other credentialing programs with which we work. The results of these additional analyses showed similar findings to those observed for the two programs that we reported above with the MAD, RMSD, and average standard deviations increasing when the number of examinees and number of exam items were fewer. In fact, the largest differences we found between the proportional curve fitting algorithm and Newton-Raphson algorithm occurred when we created missing data using the CAT algorithm with less than 30 scored item responses and fewer than 100 examinees. Although, even in these cases the differences in the item and person parameter estimates were only in the second decimal place, with the proportional curve fitting algorithm performing slightly better. These results imply that in most practical situations that choice of algorithm will not have a large impact on results and the algorithm used is a matter of preference.

We did find one situation where both algorithms faced challenges, which was when the data was not well conditioned (see Molenaar, 1995). That is, both algorithms can struggle with convergence if there are subsets of the population that respond to different items and there was not at least one item that was responded to correctly and incorrectly across populations. In Winsteps, such a situation is indicated by the message "Data are ambiguously connected" when performing the calibration. These types of situations can sometimes arise with sparse data, and the result is that item and person parameter estimates are not comparable across subpopulations.

It is also important to point out that our analyses only focused on the dichotomous Rasch model. It is possible that the algorithms may exhibit larger differences with polytomous items. For example, it is possible that one may observe larger differences when using the Rasch Partial Credit model (Masters, 1982) or the Rasch Rating Scale model (Andrich, 1978), which are designed for polytomous items. Future research should investigate the performance of proportional curve fitting and Newton-Raphson algorithms with these types of data. Future research could also look at how the convergence criterion and other settings of the algorithms may impact results. In explorations with our data, we found that the algorithms tended to be more accurate when the convergence criterion was more stringent. We also found that the number of

iterations tended to increase when the convergence criterion was more stringent, and the amount of missing data was higher. However, even in these cases both algorithms tended to quickly reach convergence; often in less than 20 iterations.

Given the stakes often attached to test scores, using algorithms that produce accurate estimates of item and person parameters are imperative. This study adds to the current literature on the how various algorithms handle sparse data by demonstrating that JMLE proportional curve fitting and Newton-Raphson algorithms often lead to very similar results when they are employed with the dichotomous Rasch model. Since different software packages often use JMLE algorithms based on one of these two approaches, researchers and practitioners should have confidence that the choice of algorithm should not lead to markedly different results.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). ACER ConQuest 3.0. [Computer software]. Melbourne: ACER.

Anderson, E. B. (1973). Conditional inference and multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *34*, 42-54.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4*, 205-221.

Bertoli-Barsotti, L., & Punzo, A. (2013). Rasch analysis for binary data with nonignorable non-responses. *Psicológica*, *34*, 97-123.

Custer, M., Sharairi, S., & Swift, D. (April, 2012). *A comparison of scoring options for omitted and not-reached item through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, British Columbia.

Del Pino, G., San Martin, E., González, J. & De Boeck, P. (2008). On the relationships between sum score based estimation and joint maximum likelihood estimation. *Psychometrika*, *13*, 145-151

DeMars, C. (April, 2002). *Missing Data and IRT item parameter estimation*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL.

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 Series B*, *39*, 1-38.

Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistical and Probability Letters*, *23*, 165-170.

Heine, J-H., & Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling, 57*, 3-36.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5*, 1148-1169.

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling, 53*, 380-393.

Jansen, P. G., van den Wollenberg, A. L., & Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12*, 297-306.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test Analysis Modules. R package version 1.995-0. [Computer software]. https://cran.r-project.org/web/packages/TAM/index.html.

Linacre, J. M. (2014). Facets: Rasch-model computer programs (Version 3.71.4) [Computer software]. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2016). Winsteps: Rasch-model computer programs (Version 3.92.0) [Computer software]. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*, 95-110.

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement*, *3*, 381-405.

Linacre, J. M. (1987). Rasch estimation: Iteration and convergence. *Rasch Measurement Transactions, 1*, 7-8.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*, 615-630.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-172.

Meyer, J. P. (2016). jMetrik (Version 4.0) [Computer software]. Charlottesville, VA: University of Virginia.

Meyer, J. P, & 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.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent development, and applications* (pp. 39-51). New York, NY: Springer-Verlag.

Nydick, S. W. (2014). catIrt: An R package for simulating IRT-based computerized adaptive Tests (Version 0.5) [Computer program]. Available at http://cran.r-project.org/web/packages/catIrt/index.html. Minneapolis, MN: Author.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Denmark: Danish Institute for Educational Research). Expanded edition with foreword and afterword by B. D. Wright,. Chicago, IL: The University of Chicago Press. Reprinted (1993) Chicago, IL: MESA Press.

Rose, N., von Davier, M., & Nagengast, B. (2010). *Modeling nonignorable missing data with IRT* (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Shin, S. (2009). How to treat omitted response in Rasch model-based equating. *Practical Assessment Research & Evaluation, 14(1).*

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing scores in test and questionnaire data. *Multivariate Behavioral Research, 38,* 505-528.

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., Gorin, J. S., Fay, D., & Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling, 55,* 335-360.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model.*Psychometrika, 58,* 395-415.

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65,* 376-404.

Willse, J. T. (2014) MixRasch: Mixture Rasch models with JMLE. R Package version 1.1. [Computer software]. https://cran.r-project.org/web/packages/mixRasch/index.html.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction. *Applied Psychological Measurement, 12,* 315-318.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37,* 573-586.

Wyse, A. E., & Babcock, B. (2016). How does calibration timing and seasonality affect item parameter estimates. *Educational and Psychological Measurement 76,* 508-527.