IRT AND SVD: IMPLEMENTING PSYCHOMETRIC METHODS IN NEW AND
COMPLEX SITUATIONS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by
Quinn N Lathrop

Ying Cheng, Director

Graduate Program in Psychology
Notre Dame, Indiana
April 2015
This document is in the public domain.
Most psychometric techniques used to analyze assessment data are designed to work with complete data. The rapid increase in the availability and power of technology has contributed to the growing use of computerized tests and related methods. The data arising from these new and complex situations challenge traditional psychometric techniques because of their size (as there is much more data) and their vast missingness (as students respond to only a small subset of possible items). This dissertation focuses on the effect of missing data on psychometric techniques.

When individuals respond to different items of varying difficulty, the psychometric techniques that rely on complete data can perform poorly. This dissertation proposes using Singular Value Decomposition (SVD), a matrix decomposition technique often seen in data mining, as a psychometric tool. The major result is that SVD is a viable psychometric tool that appears largely robust to missing data and to the missing mechanism. This document provides analytical and empirical justification for SVD’s use with psychometric data under missing data.

Chapter 1 introduces relevant IRT techniques such as nonparametric IRT and a nonparametric item fit statistic. Then, in Chapter 2, SVD is introduced and as well as an Alternating-Least-Squares (ALS) algorithm that extends the decomposition to missing data. Chapter 3 investigates the large sample properties of using SVD with
psychometric data. SVD is shown to be a consistent ordinal estimator of student ability and a consistent ordinal estimator of item easiness.

Chapter 4 presents simulation results that show that when students respond to different items of varying difficulty, whether the missingness is related to their ability or not, SVD can rank the students better than proportion correct and can better estimate the true relationship between student ability and the probability of a correct response. When missingness is related to student ability, SVD can rank the students, in most conditions, better than a parametric IRT model, even when the parametric model is correctly specified.
## CONTENTS

FIGURES ......................................................................................... iv

TABLES ........................................................................................ v

ACKNOWLEDGMENTS ..................................................................... vii

CHAPTER 1: INTRODUCTION ............................................................. 1
1.1 Nonparametric IRT ................................................................. 3
   1.1.1 Applications of NIRT .................................................... 5
   1.1.2 Ability Estimate Under NIRT ........................................ 6
   1.1.3 Rest Scores ................................................................. 7
1.2 Using NIRT to Test Parametric Item Fit ............................... 8
1.3 Missing Data and IRT ............................................................. 9
   1.3.1 Missing Data and NIRT .............................................. 12
1.4 Summary ................................................................................ 13

CHAPTER 2: SINGULAR VALUE DECOMPOSITION AS A PSYCHOMETRIC TOOL .................................................... 15
2.1 SVD under Missing Data ....................................................... 18
2.2 Demonstrations of ALS-SVD ............................................... 21
2.3 Issues of Manifest Monotonicity With SVD ......................... 24
2.4 Conclusions .......................................................................... 27

CHAPTER 3: LARGE SAMPLE PROPERTIES OF ALS-SVD IN PSYCHOMETRIC SITUATIONS .................................................. 29
3.1 Consistency of the ALS-SVD Ability Estimator ...................... 30
3.2 Consistency of the ALS-SVD Item Estimator ......................... 33
   3.2.1 Interpretation of $c$ ..................................................... 34
3.3 Discussion and Conclusions ................................................. 36

CHAPTER 4: SIMULATION STUDIES .................................................. 38
4.1 Simulated Conditions ............................................................ 38
   4.1.1 Missing Data ............................................................ 38
   4.1.2 Models and Items ....................................................... 42
   4.1.3 Generating Examinees and Item Group Membership ....... 43
FIGURES

1.1 Example of Kernel-Smoothing Procedure for a Single Item .................................................. 5
2.1 Example of ALS-SVD: Vector Construction and Matrix Multiplication .......................... 20
2.2 Demonstration of ALS-SVD Algorithm with $N = 3000$ and $I = 60$ with 50\% MCAR .................................................................................................................................................. 24
3.1 Relationship Between $c$ and $Pr(Y = 1)$ under 2PL Complete Data, MCAR, and MNAR as Person Sample Size Increases ......................................................... 37
4.1 Example of One Item from the Hyperbolic Cosine Model .................................................. 44
4.2 Strength of Spearman $\rho_s$ .................................................................................................. 47
4.3 Mean of $\rho_s$ Across All Study 1 Missing Data Conditions, 20 Responses per Person .................................................................................................................................................. 52
4.4 Study 1B: Smoothed $\rho_s$ as a Function of MNAR Correlation at Different Levels of Overlap, 20 Responses Per Person ................................................. 57
4.5 Study 1B: Smoothed $\rho_s$ as a Function of MNAR Correlation at Different Levels of Overlap, 40 Responses Per Person .................................................. 58
4.6 Study 1B: Mean $\rho_s$ as a Function of Overlap at Different Levels of MNAR Correlation, 20 Responses Per Person ......................................................... 59
4.7 Study 1B: Mean $\rho_s$ as a Function of Overlap at Different Levels of MNAR Correlation, 40 Responses Per Person ......................................................... 60

iv
TABLES

2.1 EXAMPLE OF CONVERGENCE OF ALS-SVD WITH COMPLETE DATA ................................................. 22
2.2 EXAMPLE OF CONVERGENCE OF ALS-SVD WITH MNAR DATA .................................................... 23
2.3 SNYDER’S EXAMPLE OF MANIFEST NONMONOTONICITY IN ITEM-SCORE REGRESSIONS ......................... 26
2.4 SNYDER’S EXAMPLE OF MANIFEST NONMONOTONICITY IN ITEM-SCORE REGRESSIONS WITH SVD .................. 27
4.1 MISSING DATA SIMULATION CONDITIONS ................................................................. 40
4.2 MEAN OF SPEARMAN \( \rho_s \) (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS CONDITION 1: COMPLETE DATA .................................................. 48
4.3 MEAN OF SPEARMAN \( \rho_s \) (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS CONDITION 2: 50% MCAR .............................................................. 49
4.4 MEAN OF SPEARMAN \( \rho_s \) (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS CONDITION 3: 50% MNAR .............................................................. 50
4.5 MEAN (AND STANDARD DEVIATION) OF BIAS AND ROOT-MEAN-SQUARED-ERROR OF ITEM PARAMETERS UNDER 50% MCAR AND MNAR .............................................................. 53
4.6 MEAN (AND 95% CI) OF \textit{RISE} OF ESTIMATED ICCS CONDITION 1: COMPLETE DATA .................................................. 63
4.7 MEAN (AND 95% CI) OF \textit{RISE} OF ESTIMATED ICCS CONDITION 2: 50% MCAR .............................................................. 64
4.8 MEAN (AND 95% CI) OF \textit{RISE} OF ESTIMATED ICCS CONDITION 3: 50% MNAR .............................................................. 65
4.9 GENERATING AND CANDIDATE MODEL TO TEST ITEM FIT .................................................. 67
4.10 TYPE I ERROR RATE OF \textit{RISE} FIT STATISTIC .............................................................. 71
4.11 POWER OF \textit{RISE} FIT STATISTIC .............................................................. 72
4.12 POWER OF \textit{RISE} FIT STATISTIC FOR HCM ITEMS .................................................. 73
A.1 COMPUTATIONAL TIME IN SECONDS (AND CONVERGENCE RATES) OF 2PL-IRT AND ALS-SVD FOR 1PL AND 2PL GENERATING DATA .................................................. 82
A.2 COMPUTATIONAL TIME IN SECONDS (AND CONVERGENCE RATES) OF 2PL-IRT AND ALS-SVD FOR 3PL AND HCM GENERATING DATA .................................. 83

B.1 MEAN OF SPEARMAN ρs (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 4: 75% MCAR BALANCED 86

B.2 MEAN OF SPEARMAN ρs (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 5: 75% MNAR BALANCED 86

B.3 MEAN OF SPEARMAN ρs (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 6: 75% MCAR UNBALANCED 87

B.4 MEAN OF SPEARMAN ρs (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 7: 75% MNAR UNBALANCED 87

B.5 MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 4: 75% MCAR BALANCED 88

B.6 MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 5: 75% MNAR BALANCED 88

B.7 MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 6: 75% MCAR UNBALANCED 89

B.8 MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 7: 75% MNAR UNBALANCED 89
ACKNOWLEDGMENTS

I would like to acknowledge my advisor Dr. Ying Cheng and the Quantitative Psychology faculty at the University of Notre Dame.
CHAPTER 1

INTRODUCTION

Psychological and educational human attributes, such as math ability, are often measured by assessments that consist of a series of items that the individuals respond to, as in a math test. These item responses are often theorized to come from (1) properties of the individual, (2) properties of the item, and (3) a model to link these properties to the observed response. In this paper, a person’s response to an item is scored as correct or incorrect, represented by 1 and 0 respectively. The scored response is denoted $y_{pi}$, with the subscript $p = 1, 2, \ldots, N$ for persons and $i = 1, 2, \ldots, I$ for items.

Item Response Theory (IRT) is a broad and flexible framework that attempts to models the scored response. Unidimensional models (the focus of this paper) refer to the unidimensionality of the latent ability scale; each person is represented by a single, interval-level number, denoted $\theta_p$. The unidimensional ability is the property of the person that contributes to their response. On the item side, in the simplest case, an item is represented by an item difficulty parameter $\beta_i$. Both $\theta_p$ and $\beta_i$ are on the same continuous, quantitative latent scale.

The properties of the persons and the items are linearly combined, to produce the linear component $\eta_{pi}$,

$$\eta_{pi} = \theta_p - \beta_i. \quad (1.1)$$

For binary outcome $y_{pi}$, the observed data model posits that,

$$\pi_{pi} = \frac{exp(\eta_{pi})}{1 + exp(\eta_{pi})}, \quad (1.2)$$
where $\pi_{pi}$ is the probability of a correct response for person $p$ on item $i$, and finally

$$ y_{pi} \sim Bernoulli(\pi_{pi}), $$

(1.3)

which introduces error into the system and restricts the realized outcome to 0 or 1. In other words, $Pr(Y_{pi} = 1|\theta_p, \beta_i) = \pi_{pi}$. The IRT model described above is often named the 1-parameter-logistic (1PL) model and is very similar in the mathematics to a Rasch model [Rasch, 1960].

A widely used IRT model called the 2-parameter-logistic (2PL) model adds a property to each item. The linear component of the 2PL model is calculated as,

$$ \eta_{pi} = \alpha_i \times (\theta_p - \beta_i). $$

(1.4)

The additional item parameter $\alpha_i$ is called the discrimination parameter, and weights, or scales, the ability/difficulty differences differently for each item. Item discrimination is often explained in terms of the slope of the item characteristic curves (ICC); low item discrimination results in a flatter curve, high item discrimination in a steeper curve. In this way, item discrimination can be thought of as representing the strength of the relationship between the latent ability and the observed response of the item. It has also been shown that the discrimination parameter defines the unit of the latent trait [Humphry, 2011].

In the 1PL and 2PL models, $\pi_{pi}$, the probability of a correct response for person $p$ on item $i$, is naturally bounded by 0 and 1 (as the input to the Bernoulli process is a probability). A common modification to these models involves compressing the probability scale to range between $g$ and 1, where $g \geq 0$. The lower bound $g$ is the probability that a person with $\theta_p = -\infty$ can still produce a correct response. When
added to the 2PL model, the 3-parameter-logistic (3PL) model takes the form

\begin{align}
\eta_{pi} &= \alpha_i \times (\theta_p - \beta_i) \\
\pi_{pi} &= g_i + (1 - g_i) \times \frac{\exp(\eta_{pi})}{1 + \exp(\eta_{pi})},
\end{align}

(1.5)

and so the lower asymptote \( g_i \) modifies the linking process between the linear component and the probability of correct response by compressing the probability scale from \([0, 1]\) to \([g_i, 1]\). This is theoretically meaningful because many questions, especially multiple choice questions, have a non-zero probability that a random response will be graded as correct. Because of the use of multiple choice questions, the 3PL model is arguably the most widely used model in practice. But to estimate \( g_i \), some mass of extremely low ability persons is needed. As made clear by Han (2012), “the lower asymptote of a response function begins at negative infinity on the theta scale, not at a certain point on the theta scale where low-ability examinees are observed.” It is difficult and sometimes impossible to estimate the \( g_i \) parameters (Holland, 1990). Instead, \( g_i \) can be fixed to a theoretical constant, such as one divided by the number of multiple choice options (Han, 2012). Commonly, \( g \) is called the guessing parameter.

1.1 Nonparametric IRT

The 1PL, 2PL, and 3PL IRT models belong within the family of parametric IRT models. They assume a parametric form of the ICC (logistic) that can be described by a small number of parameters. The item parameters can be combined with the parametric form of the model to calculate \( \pi_{i}(\theta) \). By plotting \( \pi_{i}(\theta) \) against \( \theta \), the ICC is formed.

Nonparametric IRT (NIRT) does not specify a parametric form of the ICC and instead attempts to estimate the probability of a correct response directly. Under NIRT, the probability of a correct response is labeled \( P_i \) and its estimate \( \hat{P}_i \). A well
known NIRT model involves the kernel smoothing procedure proposed by Ramsay (1991), and is computed as

$$\hat{P}_i(t) = \frac{\sum_p \mathcal{K}\left[\frac{\hat{\theta}_p - t}{h}\right] y_{pi}}{\sum_p \mathcal{K}\left[\frac{\hat{\theta}_p - t}{h}\right]},$$

where $\hat{\theta}_p$ is person $p$’s ability estimate, $y_{pi}$ is the observed binary response, $t$ is a given evaluation point, $\mathcal{K}$ is a symmetric, nonnegative kernel, and $h$ is the bandwidth parameter. At each given evaluation point $t$, a weighted average is taken between the observed 0’s and 1’s, with weights determined by the distance between the observed point’s ability estimate and the evaluation point. By computing this weighted average for many $t$ along the range of ability, an ICC can be graphed by plotting $\hat{P}_i(t)$ against $t$, as shown in Figure 1.1 below. Additionally, any specific value of $\hat{P}_i(t)$ can be calculated, but this calculation involves the data, and not a small number of parameters.

The bandwidth parameter $h$ dictates the trade off between bias and variance. If the bandwidth parameter is too large, the curve will be low-variance but possibly contain high bias. If the bandwidth parameter is too small, the curve will bounce between each data point resulting in a low-bias curve that chases sampling error. Thus, the choice of bandwidth parameter should balance these two forces. There are many ways to choose the bandwidth parameter, based on qualities of the data, graphical heuristics, or cross-validation (Jones et al., 1996; Ramsay, 1991). With a gaussian kernel, the optimal bandwidth parameter that minimizes mean squared error for normally distributed data can be estimated with

$$h = 1.06 \times N^{-\frac{1}{2}}$$

(Scott, 1992; Silverman, 1986).
1.1.1 Applications of NIRT

In their overview of NIRT, Junker and Sijtsma (2001) provide three broad applications of NIRT. First, NIRT can be a teaching tool that can help students understand parametric IRT models. Second, NIRT allows IRT analyses to be performed in situations where parametric models are not known to fit well, and thus NIRT can test the performance of parametric IRT models. And third, NIRT provides methods that generally work with smaller data sets (several hundred examinees) compared to those in large scale assessment (tens of thousands of examinees).

While it is true that NIRT was originally developed as an exploratory graphical procedure (Ramsay, 1991), it has also been shown that, under mild conditions, the nonparametric ICC converges to the true ICC (Douglas, 1997). Based on this early work, many procedures have been developed specifically for NIRT. One category of
procedure includes those that replace the parametric estimate of the probability of a
correct response with the NIRT estimate. A second category uses NIRT in situations
where the functional form of the ICCs is not known, and thus a parametric IRT model
may not be appropriate. NIRT assumes unidimensionality and local independence,
but makes no assumptions regarding the form of the ICCs. Thus, instead of assuming
a parametric form, NIRT can let the data decide the form the ICC. This is extremely
useful in exploratory analytics and in testing parametric item fit.

Some of the applications of NIRT beyond graphical procedures include testing
parametric model fit (Douglas and Cohen 2001; Liang and Wells 2009; Sueiro and
Abad 2011; Wells and Bolt 2008), testing parametric assumptions of unidimen-
sionality and local independence (Habing 2001), differential item function (Douglas
et al. 1996; Zheng et al. 2010), response time modeling (Ranger and Kuhn 2012),
continuous-outcome response models (Ferrando 2004), person-fit indices (Emons
et al. 2004; St-Onge et al. 2009), estimation of decision accuracy and consistency
(Lathrop and Cheng 2014), unfolding models (Johnson 2006), and as the founda-
tion for computerized adaptive testing (CAT) (Xu and Douglas 2006). The many
applications of NIRT reach almost every corner of the IRT field.

1.1.2 Ability Estimate Under NIRT

Returning to Equation 1.6, the nonparametric estimate of the probability of cor-
rect response requires \( \hat{\theta}_p \), the ability estimate for person \( p \). In most parametric IRT
procedures, the ability estimates are available only after estimating the item pa-
rameters, assuming the item parameters are known, and then estimating the ability
parameters in a separate procedure (i.e. Marginal Maximum Likelihood Estimation,
MMLE). In the nonparametric case, parametric estimates of \( \hat{\theta}_p \) are not available. In
this case, \( \hat{\theta}_p \) can be estimated by the following:

1. Calculate total score for each person \( x_p = \sum_i y_{pi} \)
2. Rank the total scores $x_p^* = rank(x_p)$, breaking ties randomly

3. Place the ranks onto the normal density as $\hat{\theta}_p = \Phi\left(\frac{x^*_p}{\sqrt{N+1}}\right)$, where $\Phi(\cdot)$ is the standard normal quantile function

The resulting $\hat{\theta}_p$ follow $\mathcal{N}(0,1)$ and are based on the ranks of the total score. Two points are worth discussion. First, the ability estimates are clearly based on ranking the persons and they are termed *ordinal ability estimates* [Douglas, 1997]. This is interesting when discussing the level of measurement of the latent scale; that is, if we are constructing a quantitative scale or not. Clearly NIRT does not, and makes the jump from ordinal to quantitative estimates as a convenience [Stout, 1990]. Any monotonic transformation of the ranked total scores will maintain the ordinal ability scale.

Second, the estimates follow $\mathcal{N}(0,1)$ in order to make the resulting ICCs comparable to parametric IRT, where (for MMLE estimation) latent ability is assumed to follow $\mathcal{N}(0,1)$, where the location and scale help solve the indeterminacy of the system. Similarly, Bayesian estimation of parametric IRT models often use $\mathcal{N}(0,1)$ as a prior for ability. When mimicking the logistic curve is not a primary goal, the nonparametric ability estimates can be placed on any scale without affecting the fit of the data [Ramsay, 1991]. Lathrop and Cheng (2014) exploit this in the situation where tests are used to make classifications and use the observed total scores in their raw form as the ability estimate. The resulting ICCs are not logistic, but the procedure allows estimates of classification accuracy and consistency to occur on the total score scale. The arbitrariness of scale is also exploited in Douglas (1997) where ability is bounded by 0 and 1 in order to facilitate certain proofs.

1.1.3 Rest Scores

There is some disagreement in the field regarding whether the nonparametric estimate of ability must first remove the item that is currently under investigation.
That is, when calculating the ICC for item \(i\), the rest score is the total score of the “rest” of the items: \(x_{p(-i)} = x_p - y_{pi}\). If not using rest scores, then the response to the item under investigation appears in both sides of the kernel smoothing equation, which violates local independence [Douglas 1997].

However, the practical implications of using or not using rest scores are not clear. As ordinal ability estimators, total score and rest score both converge to the true ranking of the examinees for monotonic, unidimensional, and locally independent items, as the number of items increases [Douglas 1997]. In the literature, a surprising number of NIRT papers do not discuss using rest scores and nonetheless produce desirable results. One downside to using rest scores is that each item is investigated on a different scale, and evaluating the response data in this piece-wise approach is at times not ideal [Lathrop and Cheng 2014; Tijmstra et al. 2011]. Also, using total scores require ranking examinees only once, where rest scores repeats the operation for each item and must store \(I\) ability estimates for each person.

1.2 Using NIRT to Test Parametric Item Fit

This section describes the item fit statistic Root-Integrated-Squared-Error, or \(RISE\), which tests parametric item fit using nonparametric ICCs. Developed by [Douglas and Cohen 2001] and evaluated by [Wells and Bolt 2008] and then expanded to the polytomous items in [Liang and Wells 2009], the \(RISE\) item fit statistic is computed by first estimating the nonparametric ICCs, \(\hat{P}_i(t)\). Recall that \(t\) are chosen evaluation points. Because the nonparametric ICC is calculated at a number of discrete points, the integration in \(RISE\) is numeric. The parametric candidate model, such as the 2PL model, is then compared to the kernel-smoothed ICC. The best fitting parametric curve that is within the specified parametric form is the one that minimizes the \(RISE\) between the parametric \(\hat{\pi}_i(t_j)\) and the nonparametric \(\hat{P}_i(t_j)\). \(RISE\) is calculated as:
\[ RISE_i = \sqrt{\frac{\sum_{j=1}^{Q} \left( \hat{\pi}_i(t_j) - \hat{P}_i(t_j) \right)^2}{Q}} \]  

where \( Q \) is the number of evaluation points. The parameters of the best fitting parametric ICC can be found by minimizing \( RISE \) with a general optimization routine.

The observed \( RISE \) will be positive, with small values indicating that the candidate model is appropriate and large values indicating misfit (that is, \( RISE \) is large if the candidate model is unable to account for the shape of the nonparametric ICC). To test the significance of \( RISE \), a bootstrap procedure is used. Under the null hypothesis that the candidate model holds, \( B \) samples are generated from \( \hat{\pi}_i(\theta) \) and \( N \theta \) values are drawn from the standard normal distribution. \( RISE \) is calculated in each of these bootstrapped samples, except when bootstrapped, it is known that the candidate model is correct. After \( B \) bootstraps, the \( B \) bootstrap replicates of \( RISE \) serve as the sampling distribution of \( RISE \) under the null hypothesis. By ranking the original observed \( RISE \) with the bootstrapped distribution, a p-value is calculated as

\[ 1 - \frac{\text{rank}(RISE)}{B + 1}, \]

so that if the observed \( RISE \) is large compared to the null distribution, it can be rejected at some nominal level, such as 0.05. Previous simulation results show that \( RISE \) controls Type I error and has higher power in all conditions compared against two other item fit statistics \( G^2 \) and \( S - X^2 \) ([Wells and Bolt](2008)).

1.3 Missing Data and IRT

Traditionally, IRT is applied to data from specific testing events where there are a large number of persons answering a relatively small number of items. Complete data occur when all individuals respond to all items. Complete data are ideal, and
many techniques in the IRT toolbox specifically require complete data. Techniques based on NIRT also require complete data in order to calculate the nonparametric estimate of ability which is based on total score.

Increasingly, assessment data are the result of computer/person interactions, and the technology used for data collection is often leveraged to produce adaptive tasks (van der Linden and Glas 2010), algorithmically generated items which can result in huge item banks (Bejar et al. 2003), and to conduct ubiquitous assessment (DiCerbo and Berhens 2014). A common consequence to all these situations is that the item bank contains vastly more items than is reasonable for a person to respond to. Further, it would be undesirable (for security and cost reasons) to present all items to all individuals.

If the item bank contains more items than is reasonable for a person to interact with, then there will be missing data; perhaps a large proportion of missing data. For example, if the item bank has 1000 items in it (which is a rather small item bank for many computer-based situations), but it is only reasonable for a person to answer 50 items (which is potentially a large testing burden), the resulting data set would be 95% missing.

While educational and psychological data are beginning to shift to this new paradigm of technologically collected data, similar data features can already be found in the field of data mining. For example, the well-known Netflix challenge, where millions of self-reported movie ratings were the source of competing predictive models, has a very similar structure to many modern data sets and what many people would describe as Big Data. First, it is immediately clear that it would be unreasonable and undesirable for every person to watch and rate every movie. Vast missingness occurs as expected (at over 98%). But the ratings are categorical, and Netflix was interested in predicting what ratings would be for unobserved person/movie combinations (so that they could make recommendations). This can be cast as an IRT problem,
in which the model predicts the probability of a rating for a given person/movie interaction.

Again, these new and complex situations, whether contextualized as a computer based mathematical test or an entertainment rating, exists in stark contrast to a traditional complete data set. Importantly, the present missingness is different from how missing data are usually discussed in statistics. Here, missing data refers to an absence of opportunity to collect a response, and not to non-response. When considering the missing mechanism, existing missing data terminology can still be used. The usual categorization as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) still apply. Generally, likelihood-based IRT inferences are valid with MCAR and MAR data but item and person parameters can be heavily biased under MNAR (Bradlow and Thomas, 1998). Classical Test Theory measures (including total score and proportion correct) are typically valid only for MCAR.

Previous work with missing data in IRT generally focuses on the situation when the test is designed to receive responses from examinees, but those responses are not given (Culbertson, 2011). Examples of situations with missing data in traditional, complete data IRT designs are found with time-constrained tests (Glas and Pimentel, 2008; Pohl et al., 2013), self-selection designs (Bradlow and Thomas, 1998), and with low-stakes testing (Rose et al., 2010).

In the context of low-stakes testing, simulations in Rose et al. (2010) show how measures such as total score, person proportion correct, and item means are biased when data was MNAR (that is, there was positive correlation between ability and response propensity). An IRT model that ignored the missing data was robust to data with 30% missingness and an ability/response propensity correlation of 0.6. Only when missingness was set to 50% and the ability/response propensity correlation to 0.8 did multidimensional IRT models, which model the response propensity
as a second dimension, show superiority over the simple IRT model. This study concludes with a statement that shows the differences in traditional and modern, computer-based assessment: “Obviously, it would be best to find a procedure for test administration that avoids nonresponse completely” (p.42).

When examinees choose the tasks they respond to from a larger item bank, Bradlow and Thomas (1998) show that simple IRT models can be heavily biased, unless the choice is unrelated to examinee ability. Their simulations present a worst-case scenario, where examinees responded to pairs of items, but only “turned-in” one of each pair, more often the one they got right. Thus, missing data was directly related to the value that was missing (MNAR). Their results may better serve as a warning of what could happen with choice-based data.

Finch (2008) also simulated MAR and MNAR data and measured the estimation of item parameters, but found that ignoring missing data performed as well as any data imputation method and so was generally recommended.

1.3.1 Missing Data and NIRT

Missing data cause a distinct problem for nonparametric techniques: ranking examinees by their total scores becomes difficult. In order for the total score ranking to hold, persons must respond to the same number of items. To demonstrate this, say that Student A responds to two items, gets both correct, and therefore has a total score of \( x_A = 2 \). Student B responds to ten items, gets only three correct, and has a total score of \( x_B = 3 \) but is ranked above Student A. This is because total score, in a practical sense, assumes all items with nonresponse are incorrect (Rose et al., 2010). The total score cannot be used to rank individuals under missing data (let alone a large proportion of missing data).

A quick fix is to use proportion correct, or total score divided by the number of
where $1()$ is an indicator function returning 1 if the condition is satisfied and zero otherwise. In this case, Student A from above would have a proportion correct $u_A = 1.0$ while Student B would have a proportion correct of $u_B = 0.3$. But proportion correct will not work when examinees respond to different sets of items that vary in difficulty. The same student could get very different proportion corrects depending on whether the set of items was “easy” or “hard.” Say Student A responds to the two easiest questions and Student B the ten hardest. Proportion correct does not account for this difference in difficulty. In addition, in a CAT situation, a goal is to provide items for each person so that their proportion correct is near a target (usually 0.5). If the CAT works perfectly, everyone will have the same proportion correct! Note that the reason proportion correct will fail to rank examinees is not dependent on the missing mechanism. Persons can respond to items of different difficulty under MCAR, MAR, and MNAR.

1.4 Summary

When educational and psychological data sets contain large amounts of missing data, existing psychometric techniques may not work as expected. Relatively simple statistics such as total score and proportion correct will not work correctly when individuals respond to different number of items or items of different difficulty. Methods that are based on these techniques are also affected, as their error will propagate to these other methods. Of particular interest to the simulation studies presented below in Chapter 4, proportion correct is the input to nonparametric kernel smoothing, and that estimated ICC is then the input to the item fit statistic $RISE$. Likelihood-based methods, like parametric IRT models, are generally robust to missing data when the
missingness is not related to ability. Even when missingness is related to ability, it seems to take a high proportion of missingness (50%) and a strong correlation between the missing indicator and ability (0.80) before the parametric IRT model becomes susceptible to meaningful error caused by missing data.

The next chapter presents a well known matrix decomposition technique that appears to have particular applicability to assessment data with missingness. The technique is able to rank examinees under missing data. In addition to providing a new way to rank examinees, the technique may also help recover the true nonparametric ICC which should lead to better performance of the RISE fit statistic.
CHAPTER 2

SINGULAR VALUE DECOMPOSITION AS A PSYCHOMETRIC TOOL

The purpose of this chapter is to propose a technique as a new ordinal estimator of ability that is applicable in missing data conditions. The technique is a matrix factorization procedure based on Singular Value Decomposition (SVD). Using matrix notation, bold-faced uppercase letters are matrixes and bold-faced lowercase letters are vectors. The $N$ by $I$ response matrix $Y$ can be decomposed into a matrix corresponding to the rows of $Y$ and a matrix corresponding to the columns of $Y$,

$$Y = RC\',$$

(2.1)

where the $R$ matrix corresponds to the rows of $Y$ and the $C$ matrix corresponds to the columns of $Y$. If $R$ and $C$ are the same rank as $Y$, then the decomposition is exact. Note that in the actual SVD, a diagonal matrix $\Sigma$ (containing the singular values) is placed in between $R$ and $C$, but not including it, its values are absorbed into the other matrixes.

If the rank of $R$ and $C$ is less than $Y$, say of rank $k$ then

$$\tilde{Y}_{(N \times I)} = R_{(N \times k)}C'_{(k \times I)}$$

(2.2)

where $\tilde{Y}$ is the best least squares approximation of $Y$ that can be achieved with a rank $k$ decomposition. That is, the values of $R$ and $C$ minimize the sum of the squared difference between the elements in $\tilde{Y}$ and the elements of $Y$. While least squares estimators have numerous beneficial properties, it is important to note that
in the psychometric situation $Y$ contains 0’s and 1’s and so will not have homoscedastic error variance. Chapters 3 and 4 investigate the consequence of using this least squares procedure on binary data, but previous research has shown that in high dimensional SVDs (where the rank $k$ is large), modifying the SVD for heteroscedasticity does not appear to improve prediction accuracy with binary and ordinal data (Moulton, 2013).

While SVD has broad uses in high dimensional space, and can even be applied as a technique to estimate the true rank of a matrix in the presence of measurement error (Kalman, 1996), the rank of the SVD used here is $k = 1$. This results in a $N \times 1$ vector $r$ that contains information about the rows (persons) and an $I \times 1$ vector $c$ that contains information about the columns (items).

A rank one SVD of binary data is similar in principal to a 1PL or Rasch IRT model where the probability of correct response is decomposed into a linear combination of the person and the item parameters (Moulton, 2013). The most important distinction between SVD and a latent trait model is that SVD models the observed data and not the probability of a correct response. SVD is closer to a Principal Component Analysis (PCA) as it does not model measurement error. This distinction is deepened when considering that both SVD (of rank one) and 1PL are unidimensional models, but the IRT’s single dimensional is latent while the SVD’s is a reorganization of the observed data. Another issue with SVD is that using a least squares criteria on binary data is inappropriate because the data are not continuous and do not have homogeneity of variance, as previously discussed.

So in comparing SVD to a latent trait model, although there are some similarities in philosophy, there are also clear differences. According to educational or psychological theory, the IRT model more closely represents both the realities of the data and the theory the explain the data. The IRT model posits that individuals and items have latent properties, and combining those properties leads to a latent probability
of correct response. The observed data are realizations of these latent probabilities. The SVD, in contrast, posits that the individuals and the items can be represented in a single observed dimension, and the observed data arises by adding the person and item properties together. The SVD model is clearly distant from educational and psychological theory, and is much less elegant than the latent trait model. To state in equations, SVD expects that \( y_{pi} = r_p \times c_i \) while the 1PL IRT model uses \( \Pr(y_{pi}) = \logit^{-1}(\theta_p - \beta_i) \).

Regardless of those distinctions, the major interest here is in using the values in the \( r \) vector to rank the individuals. With complete and balanced data, the \( r \) vector can be written as

\[
r = Yc(c'c)^{-1}
\]

so it is clear that \( r \) is a linear combination of \( Y \), with weights determined by \( c \). Compare this to proportion correct, which, by defining \( 1 \) as a \( I \times 1 \) vector of 1’s, is

\[
u = Y1(1'1)^{-1}.
\]

Proportion correct (and total score if we remove the inverted crossproduct of \( 1 \)) is also a linear combination of \( Y \), but the weights are fixed to equal 1.

Ability estimators are often structured as linear combination of the observed data. Under the 2PL model, a sufficient statistic for ability is the weighted total score with the item discriminations, \( \alpha_i \), as the item weights. Under the 1PL, where all discriminations are equal, the unweighted total score is a sufficient statistic for ability. [Ramsay (1991)] also proposes a weighted total score estimator with a two-step procedure where weights are determined by the difference between the initially estimated nonparametric ICC at the 25th and 75th ability percentiles. Thus, items that do not differentiate between the high and low examinees are down weighted. Similarly, [Monahan and Ankenmann (2010)] define and test several weighted total scores with
weights determined by the biserial correlation of the item/total score and the uni-
dimensional factor loading in an attempt to better control the Type I error of tests
of DIF. The point is, defining an ability estimator as a weighted total score, or in
this case a weighted proportion correct, is not new and has a long and successful his-
tory in the field. However, defining the item weights by SVD is a new and untested
technique.

2.1 SVD under Missing Data

Now consider the case where $Y$ contains missing data. Matrix operations like
SVD, as with most math operations, are not defined for missing data. Therefore,
an alternative implementation is required. To ease the use of notation, define $y_p$ as
the vector of observed responses for person $p$. Note that any missing data is not
represented or stored in this vector. Then the proportion correct for person $p$ is

$$u_p = y_p 1(1'1)^{-1}$$  \hspace{1cm} (2.3)

where 1 is a vector of 1’s of appropriate length.

For SVD, the $r$ and $c$ vectors can be found with Alternating Least Squares (ALS).
Using the ALS implementation of SVD on complete data is computationally slower,
as there are efficient algorithms to compute the decomposition directly, but ALS-SVD
allows for arbitrary missingness. With ALS-SVD under complete data, the following
two equations are iteratively computed until convergence (the $c$ vector begins with
some given initial values):

$$r = Yc(c'c)^{-1}$$

$$c' = (r'r)^{-1}r'Y$$

Under missing data, each of the above equations must be computed separately for
each element in the vector left of the equals sign. Recall that $y_p$ is the row vector of observed responses for person $p$ and further define $c_p$ to be the elements of $c$ that correspond to the elements of $y_p$. For example, if $y_p$ contains person $p$’s responses to items 1, 3, and 4 (see Figure 2.1 below for a demonstration), then $c_p$ contains the item weights for items 1, 3, and 4 as well. Each element of the first equation is calculated as

$$r_p = y_p c_p (c_p' c_p)^{-1}.$$ 

Note that when updating the elements in $r$, the elements of $c$ are held fixed. After all elements in the $r$ vector are updated, then the elements of the $c$ vector are updated. To do so, define the vector $y_i$ as the column vector of observed responses for item $i$ and $r_i$ to be the fixed elements of $r$ that correspond to the elements of $y_i$.

$$c_i' = (r_i' r_i)^{-1} r_i' y_i.$$ 

Convergence occurs when the change in the elements of $r$ and $c$ become arbitrarily small from iteration to iteration. After convergence, which requires generally a small number of iterations, the resulting score for person $p$ is,

$$r_p = y_p c_p (c_p' c_p)^{-1}$$

which is again a linear combination of $y_p$ with weights given by $c_p$.

There has been one previous line of research combining SVD and IRT, although with different motivations. [Moulton] (2013) provides an ALS-SVD algorithm to use SVD to impose IRT-like measurement on big data sets. Notably, the main application of the work is with highly dimensional response data. The algorithm developed is called Non-Unidimensional Scaling (NOUS). The distinct feature of the present work compared to [Moulton] (2013), is that the present work provides new implementations
of existing methods while Moulton (2013) provides an entirely new framework with which to analyze data. For example, new definitions are given for Standard Error, Stability, Objectivity, and other terms, all of which are functions of the multidimensional ALS-SVD and its results. The joining of SVD and IRT in the present paper does not try to develop a new framework, but instead, argues that the unidimensional ALS-SVD studied here belongs in line with many traditional IRT ability estimators.

\[
\begin{align*}
Y &= \\
\begin{bmatrix}
0 & 1 & 0 & 1 \\
NA & NA & 1 & 0 \\
0 & 1 & 1 & NA \\
1 & 1 & 1 & 0 \\
1 & 0 & NA & 1 \\
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
C &= \\
\begin{bmatrix}
.50 \\
.67 \\
.75 \\
.50 \\
.50 \\
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
r_p &= Y_p \\
Y_p &= \begin{bmatrix}
0 & 0 & 1 \\
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
C_p &= \begin{bmatrix}
.50 \\
.75 \\
.50 \\
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
l_{-1} &= \begin{bmatrix}
.50 \\
.75 \\
.50 \\
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
C_{l_{-1}} &= \begin{bmatrix}
.50 \\
.75 \\
.50 \\
\end{bmatrix}
\end{align*}
\]

---

**Figure 2.1.** Example of ALS-SVD: Vector Construction and Matrix Multiplication
Figure 2.1 shows a small example of how the mechanics of ALS-SVD work. Here, \( Y \) is a 5 \( \times \) 4 matrix with missing values coded as NA. To begin the ALS-SVD, \( c \) is filled with initial values (here, the pairwise deleted column means). Then, each element of \( r \) is computed one at a time. The first row of \( Y \) corresponds to the first element of \( r \). To compute \( r_p \) for the first person, two vectors are first formed, \( y_p \) and \( c_p \). Taking the first row of \( Y \) and removing the missing elements leads to the \( y_p \) shown in the lower portion of Figure 2.1. Similarly, \( c_p \) contains the values of the \( c \) vector that correspond to the elements of \( y_p \). Because the element in the first row second column of \( Y \) is missing, \( y_p \) does not contain this cell and \( c_p \) also does not contain the second element of the \( c \) vector. The result of the matrix multiplication is 0.47, which is the first element of \( r \). In turn, all the elements of \( r \) are computed. Then, the updated \( r \) vector is used to update the elements of the \( c \) vector. This alternates until convergence.

See Appendix A for R code that implements ALS-SVD for person by item response matrixes. Appendix A also contains discussion on the computation features of ALS-SVD and simulation results regarding convergence and computation time of the algorithm.

2.2 Demonstrations of ALS-SVD

This section demonstrates how ALS-SVD reaches convergence in its iterations. Chapter 3 goes into detail about the true values that ALS-SVD converges to, but for now, the claim is that at convergence, \( r \) is an ordinal estimator of \( \theta \) and \( c \) is an ordinal estimator of \( Pr(Y = 1) \), the probability of a correct response integrated over the \( \theta \) distribution. Chapter 3 will provide evidence to support these claims.

For the first demonstration, assume that \( \theta \) takes only two district values, \( \theta_1 < \theta_2 \), both occurring at equal probability in the population. Say there are three items, for which the probability of correct response for examinees in the first group are
\(\pi_1(\theta_1) = 0.8, \pi_2(\theta_1) = 0.5, \text{ and } \pi_3(\theta_1) = 0.2\) and for the second group \(\pi_1(\theta_2) = 0.9, \pi_2(\theta_2) = 0.8, \text{ and } \pi_3(\theta_2) = 0.5\). If all examinees respond to all items, or if any missingness is completely at random, then ALS-SVD converges in the population to the solution shown in Table 2.1 (note the initial values for \(c^0\) are 1, 1, and 1).

### Table 2.1

**Example of Convergence of ALS-SVD with Complete Data**

<table>
<thead>
<tr>
<th></th>
<th>(c^0)</th>
<th>(r^1)</th>
<th>(c^1)</th>
<th>(r^2)</th>
<th>(c^2)</th>
<th>(r^3)</th>
<th>(c^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.500</td>
<td>1.346</td>
<td>0.525</td>
<td>1.350</td>
<td>0.525</td>
<td>1.350</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.733</td>
<td>1.062</td>
<td>0.716</td>
<td>1.059</td>
<td>0.716</td>
<td>1.059</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.592</td>
<td>0.587</td>
<td>0.587</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

At convergence, \(r_1 = 0.525\) and \(r_2 = 0.716\) and the three items are correctly ordered in terms of easiness. Note particularly how quickly ALS-SVD converges from the initial values (which correspond to the proportion correct estimator). After the first iteration, the change is minimal and the algorithm converges after only three iterations.

With the same conditions as above, now say that all examinees with \(\theta_1\) do not respond to item 3 (the hardest item) and that all examinees with \(\theta_2\) do not respond to item 1 (the easiest item). Both ability groups still respond to item 2. This is a simple case of MNAR, as the missingness is due to ability. The convergence of ALS-SVD
in this case is shown in Table 2.2 below. How does ALS-SVD cope with a situation where examinees of different abilities respond to items with different probabilities of correct response? Using the same initial values for $c^0$, the values for $r^1$ suggest that $\theta_1 = \theta_2$ because if all items are weighted equally, as they are with proportion correct, the average probability of a correct response to any item that an examinee responds to is the same for both ability groups. In the next step of the algorithm, $c^1$ is calculated from the values in $r^1$. The items are then ordered in terms of their easiness and in the next calculation, $r^2$ correctly orders the ability groups. With the MNAR missing data, the algorithm requires more iterations to converge, but does so after 11 iterations. Note again, however, how quickly the algorithm moves from its starting values so that after only a few iterations, the movement in $r$ and $c$ is small.

To view the convergence of ALS-SVD on a larger scale, consider the following demonstration data shown in Figure 2.2. This example data has 3,000 examinees responding to 60 2PL items. The data contains 50% MCAR missing data, which

<table>
<thead>
<tr>
<th>$c^0$</th>
<th>$r^1$</th>
<th>$c^1$</th>
<th>$r^2$</th>
<th>$c^2$</th>
<th>$r^3$</th>
<th>$c^3$</th>
<th>$r^4$</th>
<th>$c^{10}$</th>
<th>$r^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.650</td>
<td>1.231</td>
<td>0.590</td>
<td>1.355</td>
<td>0.561</td>
<td>1.425</td>
<td>0.546</td>
<td>1.516</td>
<td>0.527</td>
</tr>
<tr>
<td>1</td>
<td>0.650</td>
<td>1.000</td>
<td>0.744</td>
<td>0.987</td>
<td>0.790</td>
<td>0.972</td>
<td>0.813</td>
<td>0.949</td>
<td>0.843</td>
</tr>
<tr>
<td>1</td>
<td>0.769</td>
<td>0.672</td>
<td>0.633</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.593</td>
<td></td>
</tr>
</tbody>
</table>
slows down the convergence of ALS-SVD so the iterations are more visually distinct. As before, the initial values used for $c$ are all equal to 1 (the proportion correct estimator). After the first estimation of $r$, which takes only the discrete values allowed by proportion correct, the item easiness is estimated. By the fifth iteration, the movement in $r$ and $c$ is small, and both show a monotonic relationship with their true baseline ($Pr(Y = 1)$ for $c$ and $\theta_p$ for $r$, see Chapter 3 for details). The algorithm converges in the tenth iteration. Again, the major movement of the ALS-SVD algorithm occurs in the first few iterations.

### 2.3 Issues of Manifest Monotonicity With SVD

For the traditional use of NIRT with complete data, total score and nonparametric ICCs will jointly converge to their true values (Douglas, 1997). There are situations, however, where, even when the true ICCs are monotonic, the total scores will lead to nonmonotonic item-score regressions (Junker and Sijtsma, 2001). An item-score
regression is the average of correct responses across all examinees with a particular observed score. We intuitively expect that when there is latent monotonicity in the item-trait regression (the true ICC) that should lead to manifest monotonicity in the item-score regression. The item-score regressions can be thought of as a simplified case of the kernel-smoothed ICCs, with the goal of estimating the probability of correct response as function of the observed total scores.

Snyder’s famous example of this, as presented in Snyder (2001), sets up a simple three item situation. Say that ability only takes on two values, $\theta_1$ and $\theta_2$, both occurring at 50% in the population of examinees and with $\theta_1 < \theta_2$. There are three items in the test and for the first ability level $\pi_{1,2,3}(\theta_1) = \gamma$ and for the second ability level $\pi_1(\theta_2) = 0.5$ and $\pi_{2,3}(\theta_2) = 1 - \gamma$. Allow $\gamma$ to be a number less than 0.5. All the items have an increasing $\pi$ in $\theta$, and therefore have the property of latent monotonicity. However, as $\gamma$ approaches 0 the probabilities of a correct response for item 1 for total scores $x = 0, 1, 2, 3$ are 0, .25, 0, and 1. Therefore, manifest monotonicity does not hold with the total scores for item 1. If instead we use the rest scores, item 1 remains monotonic. With rest scores of $x_{(i)} = 0, 1, 2$, the probabilities of correct response for item 1 are 0, .25, .50, which maintain manifest monotonicity. These results are displayed in Table 2.3 below.

When using SVD to rank the examinees and examining item $i$, should item $i$ be removed from the SVD ability estimation? The analogous rest score would be

$$r_{p(-i)} = y_{p(-i)}c_{p(-i)}(c'_{p(-i)}c_{p(-i)})^{-1}$$

which removes item $i$ from the response vector and the item weights. Thus the ranking is not based on the item under investigation (although that item did play some role in estimating $c$).

With SVD, each response pattern will receive a unique value in $r$, thus, the
Table 2.3

Snyder’s Example of Manifest Nonmonotonicity in Item-Score Regressions

<table>
<thead>
<tr>
<th>Total Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y_1 = 1</td>
<td>x)$</td>
<td>0</td>
<td>.25</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rest Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y_1 = 1</td>
<td>x_{(-i)})$</td>
<td>0</td>
<td>.25</td>
</tr>
</tbody>
</table>

The monotonicity of the item-score regressions clearly breaks down, because for any value in $r$, the mean of correct/incorrect responses will always be 0 or 1. The item-score regressions based on $r$ will bounce between 0 and 1 but never take any other value.

Thus, using the rest-SVD rankings $r_{(-i)}$ appears critically important. If we apply SVD to Synder’s example, we see this issue exactly. Table 2.4 shows the item-score regression for $r$ and $r_{(-i)}$. Using $r$, the item-score regression, as expected, bounce from 0 to 1 and nowhere in between. In contrast, using $r_{(-i)}$ maintains manifest monotonicity. That is because the item-$r_{(-i)}$ regression asks what is the mean of correct/incorrect responses matching examinees on their response pattern to all other items. Therefore, values other than 0 and 1 are possible.

Even though it appears $r_{(-i)}$ is preferred when examining a particular item, in Chapter 4, both $r$ and $r_{(-i)}$ are tested in estimating nonparametric ICCs. Understanding the practical implications of using $r$ and $r_{(-i)}$ is important because $r_{(-i)}$ is computationally more demanding, requires more data to be stored, and places examinees on different scales for each item.
TABLE 2.4

SNYDER’S EXAMPLE OF MANIFEST NONMONOTONICITY IN ITEM-SCORE REGRESSIONS WITH SVD

<table>
<thead>
<tr>
<th>$r$</th>
<th>0</th>
<th>.36</th>
<th>.65</th>
<th>1.02</th>
<th>1.31</th>
<th>1.68</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y_1 = 1</td>
<td>x)$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$r_{(-i)}$</th>
<th>0</th>
<th>.76</th>
<th>1.52</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y_1 = 1</td>
<td>u)$</td>
<td>0</td>
<td>.25</td>
</tr>
</tbody>
</table>

2.4 Conclusions

There is a paradigm shift occurring in educational and psychological data that implies that future data sets will often contain high percentages of missing data. A variety of psychometric techniques will not perform well under missing data, such as ranking individuals on their observed proportion correct as well as the entire family of nonparametric IRT models and the many related techniques like the item fit statistic $RISE$. In order to ask the same psychometric questions with this new and complex data, advances to foundational psychometric techniques are needed.

ALS-SVD provides a new way to rank examinees that is computationally robust to any amount or pattern of missing data. Thus, ALS-SVD may provide answers to common psychometric question when current techniques fail. ALS-SVD can be thought of as a weighted proportion correct, and therefore follows a long tradition of ability estimators in IRT. With ALS-SVD, the weights used in the ability estimate are also computed from the ALS-SVD. In Chapter 3, the large-sample properties of ALS-SVD are investigated analytically. In Chapter 4, simulation studies empirically quantify the benefit of using ALS-SVD in missing data situations over proportion correct and traditional parametric IRT models. How this benefit propagates to tech-
niques such as NIRT and *RISE* is also evaluated.
CHAPTER 3

LARGE SAMPLE PROPERTIES OF ALS-SVD IN PSYCHOMETRIC SITUATIONS

Large-sample properties of ranking examinees based on linear combinations of the observed responses has a strong history in IRT literature [Douglas 1997, Stout 1990]. Under often-used assumptions about the response process, this chapter investigates similar large-sample properties of the ALS-SVD procedure, with particularly focus on using ALS-SVD as an ordinal ability estimator.

The justification for an ordinal latent scale can be found in Stout (1990), where the author argues that unless the $\theta$ scale has been previously established and its quantitative values have specific meaning, any scale is equally valid for any strictly increasing transformation. Most often, the scale of $\theta$ is chosen for convenience. This, Stout (1990) argues, means that inferences only depend on the ordinal nature of the chosen scale. This same argument is used here in evaluating the ordinal estimators below.

Regarding notation, this chapter does not use bold-face lettering for matrices and vectors, but does employ a random variable and realization notation, where $Y$ is the random variable and $y$ its realization. Three core assumptions are used throughout this chapter:

A1 The latent ability $\theta$ is unidimensional, continuous, and scalar.

A2 Item responses are independent conditional on $\theta$, $Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_I = y_I|\theta) = \prod_{i=1}^{I} Pr(Y_i = y_i|\theta)$. This condition is called local independence.

A3 The ICCs of all items are monotonic in $\theta$, $P_i(\theta_j) \leq P_i(\theta_k)$ for all $i = 1, 2, \ldots, I$ where $\theta_j < \theta_k$. 

29
These assumptions are common assumptions in IRT, and have been referred to as a weak set of assumptions as there is no particular parametric form of the ICCs (Wells and Bolt, 2008). Assumption A3 only makes a monotonic statement, and makes no further restriction about the shape or parametric form of the ICCs, or relationships among different items’ ICCs, beyond that defined by assumptions A1 and A2. This chapter uses $P_i(\theta)$ to represent the ICC without the distinction of being parametric or nonparametric. The following does not require knowledge of form of the ICC, just that it exists and follows the above assumptions. That the ICC can be modeled with a certain parametric form or not is not of concern.

3.1 Consistency of the ALS-SVD Ability Estimator

This section establishes consistency of estimating ability on an ordinal scale with $r_p = y_p c_p (c'_p c_p)^{-1}$. For a single individual, dropping the $p$ person subscript and written not in matrix form, the estimator is,

$$r = \sum_{i=1}^{I} c_i Y_i \over \sum_{i=1}^{I} c_i^2$$

The restriction on $c$ is that $0 < c_i < \infty$. In the special case where all $c_i = 1$, the proportion correct estimator, $0 \leq r \leq 1$ and in any other case, the upper bound of $r$ depends on $c$. The scale is not of direct concern, however, because of the ordinal nature in which $\theta$ is being estimated (Stout, 1990).

The monotonicity of the true-ALS-SVD estimator can be examined by replacing $Y_i$ with the expected values $P_i(\theta)$. The true-ALS-SVD estimator is then

$$r^*(\theta) = \sum_{i=1}^{I} c_i P_i(\theta) \over \sum_{i=1}^{I} c_i^2$$

Clearly, $\sum_{i=1}^{I} P_i(\theta)$ is nondecreasing in $\theta$ under assumptions A1, A2, and A3. Because
of the restrictions on $c$, and that the transformation of $P_i(\theta)$ is without regard to the value of $\theta$, the true-ALS-SVD estimator is also nondecreasing in $\theta$.

While assumption A3 states that each item’s ICC is non-decreasing in $\theta$, it is possible to weaken this assumption to state that $\sum_{i=1}^{I} P_i(\theta)$ is nondecreasing in $\theta$. This condition is called Weakly Monotone by [Stout 1990] and allows the test to contain a relatively small number of nonmonotonic items. The following continues to use assumption A3, although in most cases, the Weakly Monotone assumption could be used with little modification.

To examine if the ALS-SVD estimator converges to the true-ALS-SVD estimator as test length grows, it is necessary to embed a finite test in an infinite collection of items [Douglas 1997]. This has been justified in previous work as analogous to creating a test from an infinitely large item bank, which is a similar set up to many computer-based testing programs [Stout 1990]. For this, a triangular set of tests of growing length is used. Let $Y^{(I)}$ (with no item subscript) denote the test of length $I$. The triangular sequence of tests is

$$
Y^{(1)} = (Y_1^{(1)}),
Y^{(2)} = (Y_1^{(2)}, Y_2^{(2)}),
\vdots
Y^{(I)} = (Y_1^{(I)}, Y_2^{(I)}, \ldots, Y_I^{(I)}).
$$

There is no requirement that subsequent tests overlap in their items. Assumptions A1, A2, and A3 hold across all test (that is, all tests measure the same latent trait in assumption A1). Also define $c_i^{(I)}$ with $i = 1, 2, \ldots, I$ as the fixed values of $c$ corresponding to the items in $Y^{(I)}$. For all $i$ and $I$, $0 < c_i^{(I)} < \infty$. Regarding all $c_i^{(I)}$ being larger than 0, the only way for a $c_i$ to equal exactly 0 is if that item is responded to incorrectly by all examinees. In the population, this implies an
infinitely difficult item. In a sample, this item would contain no variance, and thus contribute no information about the examinees. Therefore it is reasonable to exclude these items from the following sections. Another way to understand this restriction is require $P_i(\theta) > 0$ for some $\theta$. To simplify notation, define the scoring scheme $w_i = c_i / \sum_{i=1}^{I} c_i^2$ so that $r = \sum_{i=1}^{I} w_i y_i$ where $w_i$ represents the item weight.

**Theorem 3.1.1** Conditional on $\theta$, if assumptions A1, A2, and A3 hold and $0 < c_i^{(I)} < \infty$ for all $i$ and $I$, the ALS-SVD estimator approaches its true value in probability. That is, for any given $\epsilon > 0$,

$$
\lim_{I \to \infty} Pr \left( \left| \sum_i w_i^{(I)} y_i^{(I)} - \sum_i w_i^{(I)} P_i^{(I)}(\theta) \right| > \epsilon | \theta \right) = 0.
$$

**Proof.** By Chebyshev’s inequality,

$$
Pr \left( \left| \sum_i w_i^{(I)} y_i^{(I)} - \sum_i w_i^{(I)} P_i^{(I)}(\theta) \right| > \epsilon | \theta \right) \leq \frac{Var \left( \sum_i w_i^{(I)} y_i^{(I)} | \theta \right)}{\epsilon^2}
$$

$$
= \frac{1}{\epsilon^2} \sum_i w_i^{2(I)} P_i^{(I)}(\theta)(1 - P_i^{(I)}(\theta))
$$

$$
\leq \frac{\sum_i \left( \frac{c_i^{(I)}}{\sum_i c_i^{(I)}} \right)^2}{4 \epsilon^2}
$$

$$
= \frac{1}{4 \epsilon^2 \sum_i c_i^{2(I)}} \xrightarrow{I \to \infty} 0,
$$

because $P(\theta)(1 - P(\theta)) \leq 1/4$ and $\sum c_i^{2(I)}$ increases as $I$ increases. \qed

In the special case when all $c_i = 1$ (the proportion correct estimator), then the above results are similar to those in Johnson (2006) and Junker (1991) regarding the consistency of particular bounded scoring schemes (such as weighted total score).

Note that so far, the only restrictions on $c_i$ is that they are positive and finite. With any such $c_i$, the estimator is a consistent ordinal estimator of $\theta$. The next
section looks at the properties of estimating $c_i$ with ALS-SVD.

3.2 Consistency of the ALS-SVD Item Estimator

This section investigates the second half of the ALS-SVD process, estimating $c_i$ while holding $r$ fixed. This section proceeds similarly to the previous section, except that it concerns the estimation of $c$ for a single item, as the number of examinees increases. The estimator in question is,

$$c = \frac{\sum_{p=1}^{N} r_p Y_p}{\sum_{p=1}^{N} r_p^2}.$$ 

The restriction on $r$ are that $0 < r_p < \infty$ for all $p$. Because $r_p$ can equal 0 if that person’s response vector is all 0’s, it is clear that the estimator for $c$ is undefined if all $r_p = 0$, though this would only result in the impractical situation that the entire response matrix is all 0’s. As in the previous section, it is reasonable to require that all $r_p$ are larger than 0, as $r_p$ that is equal to 0 implies $P_i(\theta_p) = 0$ for all $i$.

Similar to the triangular set of tests, here a triangular set of examinees is used. Let $\Theta^{(N)}$ (with no person subscript) denote a group of examinees with $N$ individuals that respond to a single (given) item. Then

$$\Theta^{(1)} = (\Theta_1^{(1)}),$$

$$\Theta^{(2)} = (\Theta_1^{(2)}, \Theta_2^{(2)}),$$

$$\vdots$$

$$\Theta^{(N)} = (\Theta_1^{(N)}, \Theta_2^{(N)}, \ldots, \Theta_N^{(N)}).$$

Subsequent groups of examinees are not required to be subsets of any previous group, but note that in all groups assumption A1 holds (that is, $\Theta$ references the same latent trait in all groups). For every group of examinees, each examinee has associated $\theta_p$. 

33
and \( r_p \) values. Thus, for each group, there are vectors \( \theta_p^{(N)} \) of person latent abilities, \( r_p^{(N)} \), and \( Y_p^{(N)} \) of item responses, for \( p = 1, 2, \ldots, N \). Every group of examinees is drawn from an infinite population of examinees with population density \( g(\theta) \). To simplify notation, define the scoring scheme \( v_p = r_p / \sum_{p=1}^{N} r_p^2 \) so that \( c = \sum_{p=1}^{N} v_p y_p \).

**Theorem 3.2.1** Given the conditions of Theorem 3.1.1 and \( 0 < r_p^{(N)} < \infty \), the ALS-SVD estimator of \( c \) approaches its true value in probability. That is, for any \( \epsilon > 0 \),

\[
\lim_{N \to \infty} Pr \left( \left| \sum_p v_p^{(N)} y_p^{(N)} - \sum_p v_p^{(N)} P(\theta_p^{(N)}) \right| > \epsilon \left| \theta_p^{(N)} \right|_{p=1,2,\ldots,N} \right) = 0.
\]

**Proof.** Follows that of Theorem 3.1.1

\[
Pr \left( \left| \sum_p v_p^{(N)} y_p^{(N)} - \sum_p v_p^{(N)} P(\theta_p^{(N)}) \right| > \epsilon \left| \theta_p^{(N)} \right|_{p=1,2,\ldots,N} \right) \leq \frac{Var \left( \sum_p v_p^{(N)} y_p^{(N)} | \theta_p^{(N)} \right)_{p=1,2,\ldots,N}}{\epsilon^2} \leq \frac{1}{\epsilon^2} \sum_p v_p^{2(N)} P(\theta_p^{(N)})(1 - P(\theta_p^{(N)})) \leq \frac{1}{4\epsilon^2} \frac{\left( \sum_p r_p^{(N)} \right)^2}{\sum_p r_p^{2(N)}},
\]

\[
= \frac{1}{4\epsilon^2} \sum_p r_p^{2(N)} \to 0 \quad N \to \infty.
\]

\( \square \)

### 3.2.1 Interpretation of \( c \)

It is not yet clear what the true value of \( c \) means. The use of \( r \) was motivated as an estimation of the rank-order of \( \theta \). But when considering the items, what is the true value of \( c \) that the estimator in Theorem 3.2.1 converges to?
The true value of the estimator, by replacing $Y_p$ with its expected value $P(\theta_p)$, is

$$c^* = \frac{\sum_{p=1}^{N} r_p P(\theta_p)}{\sum_{p=1}^{N} r_p^2}.$$  

So the true value $c^*$ is a weighted sum of the probability of correct response to the item. To understand the role of the true-ALS-SVD $c^*$ estimator, consider that, holding $r$ constant (so the same group of individuals responds to multiple items), as $\sum P(\theta_p)$ increases, so does $c$. Therefore, $c$ is, in loose terms, an “easiness” parameter, in that a larger $c$ corresponds to a higher probability of a correct response across the examinees. The difficulty parameter $\beta$ in the 1PL and 2PL model is defined as the value of $\theta$ where $P(\theta) = 0.5$. In contrast, $c$ does not represent a single point on the ability scale that has a particular psychometric meaning, but instead relates to the easiness of the item across $r$.

Consider that as a triangular set of randomly drawn examinees increases $\frac{1}{N} \sum_p Y_p \rightarrow \frac{1}{N} \sum_p P(\theta) \approx \int P(\theta) g(\theta) \, d\theta = Pr(Y = 1)$. The true-ALS-SVD $c^*$ estimator is a monotonic transformation of the probability of correct response integrated over the population of examinees. The monotonic transformation is defined by the scale of $r$. Interestingly, both $c$ and $Pr(Y = 1)$ share a common point regardless of the scale of $r$. If for an item $P(\theta) = 0$ for all $\theta$, then both $c$ and $Pr(Y = 1)$ will equal 0. At the other extreme $P(\theta) = 1$ for all $\theta$, $Pr(Y = 1) = 1$ and $c = \sum r / \sum r^2$.

To see this behavior in example data, Figure 3.1 displays the results of a demonstrative simulation. Each panel involves the same $I = 100$ 2PL test. From left to right, the person sample size increases from 100, 1,000, 10,000, to 100,000. The horizontal axis is the true integral of the probability of correct response over the true ability distribution. The top row is complete data, the middle row 50% MCAR, and the bottom row 50% MNAR. The details of the simulation are described in Chapter 4 where these conditions correspond to Conditions 1, 2, and 3 with the difference
that in this case $I = 100$. The vertical axis in each panel is the value of $c$ after the ALS-SVD converged.

As the number of examinees increases, the relationship between $c$ and $\int P(\theta)g(\theta)\,d\theta$ gets stronger. The relationship in strongest in the complete data condition and is markedly worse, but still present, in the MNAR condition.

3.3 Discussion and Conclusions

As SVD is a procedure intended for continuous data, and has not been previously investigated on unidimensional, binary psychometric data, the above results show that the properties of ALS-SVD are nonetheless psychometrically desirable. In particular, the ALS-SVD algorithm provides a consistent ordinal estimator of ability and a consistent ordinal estimator of item easiness. Moreover, the properties of ALS-SVD are psychometrically meaningful. The meaning of an ordinal ability estimate is clear and widely used in the psychometric field. The meaning of the easiness of an item is shown to be related to the probability of correct response integrated over the ability distribution. This relationship connects the familiar $\theta$ scale and $P(\theta)$ to the ALS-SVD output.

Previously it was argued that the ALS-SVD ability estimator can be thought of as a linear combination of the observed responses, and that many other similar ability estimators exist. The weights of the linear combination defined by ALS-SVD now have a meaning that ties them back to the probability of correct response and the original $\theta$ scale.
Figure 3.1: Relationship Between $c$ and $Pr(Y = 1)$ under 2PL Complete Data, MCAR, and MNAR as Person Sample Size Increases
CHAPTER 4

SIMULATION STUDIES

This chapter describes simulation studies that evaluate the performance of ALS-SVD in a variety of situations. There are three distinct but related simulation studies in this chapter. In Study 1 (and a follow up study named Study 1B), ALS-SVD is compared against proportion correct and a parametric measure of ability in terms of ranking the examinees. In Study 2, the ranking estimators in Study 1 are used to construct nonparametric ICCs, and those ICCs are compared to a baseline derived from the true item parameters and that of a parametric IRT model. In Study 3, the nonparametric ICCs from Study 2 are used to test item fit. Error in the original examinee rankings can propagate to the nonparametric ICCs, whose error in turn can propagate to the item fit statistic. All the simulations use a similar set of conditions, which are described in the following section.

4.1 Simulated Conditions

In the simulation studies, several factors are manipulated to generate data with specific features. The simulations control the type and amount of missing data, the item bank size, the person sample size, the generating response model, and the statistical properties of the items.

4.1.1 Missing Data

The major factor of focus in the simulation studies is missing data. There are three amounts of missingness: complete data, 50% missing, or 75% missing. The
percentage of missingness refers to the overall missingness of the response matrix across all examinees and items. The complete data condition has no missing data, and so serves as a baseline to compare the results to previous studies and traditional assessment situations. The 50% missing rate is at the high end of what has been simulated in previous studies (Glas and Pimentel 2006; Rose et al. 2010). The 75% missing rate has half as much data as the 50%, but, as previously discussed, it is still optimistic compared to missingness above, say, 95%.

Changing the amount of missing data necessarily introduces imbalance in the person and item sample sizes. The number of items and persons are manipulated in such a way as to create meaningful comparisons across and within different missing data conditions. Specifically, the amount of data per person and per item is held constant across many conditions. So as the percentage of missing data increases, the size of the response matrix also increases, but the amount of information per person and per item stays the same. Table 4.1 displays the missing data conditions.
### Table 4.1

**Missing Data Simulation Conditions**

<table>
<thead>
<tr>
<th>Number</th>
<th>Condition Name</th>
<th>Item Group</th>
<th>Item Selection</th>
<th>Data Per Person</th>
<th>Data Per Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Complete</td>
<td>20 (or 40)</td>
<td>None</td>
<td>20 (or 40)</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>50% MCAR</td>
<td>40 (or 80)</td>
<td>Random</td>
<td>20 (or 40)</td>
<td>500 and 1000</td>
</tr>
<tr>
<td>3</td>
<td>50% MNAR</td>
<td>40 (or 80)</td>
<td>Ability-based</td>
<td>20 (or 40)</td>
<td>500 and 1000</td>
</tr>
<tr>
<td>4</td>
<td>75% MCAR Balanced</td>
<td>80 (or 160)</td>
<td>Random</td>
<td>20 (or 40)</td>
<td>≈ 500 and 1000</td>
</tr>
<tr>
<td>5</td>
<td>75% MNAR Balanced</td>
<td>80 (or 160)</td>
<td>Ability-based</td>
<td>20 (or 40)</td>
<td>≈ 500 and 1000</td>
</tr>
<tr>
<td>6</td>
<td>75% MCAR Unbalanced</td>
<td>80</td>
<td>Random</td>
<td>10 to 30</td>
<td>≈ 500 and 1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>Random</td>
<td>20 to 60</td>
<td>≈ 500 and 1000</td>
</tr>
<tr>
<td>7</td>
<td>75% MNAR Unbalanced</td>
<td>80</td>
<td>Ability-based</td>
<td>10 to 30</td>
<td>≈ 500 and 1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>Ability-based</td>
<td>20 to 60</td>
<td>≈ 500 and 1000</td>
</tr>
</tbody>
</table>
Starting with the complete data condition (Condition 1), the number of examinees $N$ is set to 1000 and the number of items $I$ set to 20 or 40. The size of this response matrix is considered a normal, reasonable size in the IRT literature. Person sample sizes less than 500 are considered small, while sizes of 5000 or larger are considered big. Similarly, a 20 and 40 item test would be considered a medium and long length, respectively. Although not usually discussed with complete data, from each person there are 20 or 40 pieces of data and from each item 1000 pieces of data.

Missing data patterns in Conditions 2 through 7 are generated by a two stage sampling design. First, a person is placed into one of three item groups. These can be thought of as classrooms, chapters within a book, or different forms of a test. The item groups are ordinal and are labeled Beginning, Intermediate, and Advanced. Each item group contains $I/2$ items and the Intermediate item group contains a random half of the Beginning and the Advanced item groups. As there is no direct overlap between the Beginning and Advanced groups, the Intermediate group serves to anchor both to a common scale.

In the 50% missing conditions (Conditions 2 and 3), a person responds to all items within their item group. That means each person responds to $I/2$ items. Because $I$ in Conditions 2 and 3 is set to either 40 or 80, that means that each person still responds to either 20 or 40 items, so that the inferences made about each person are made with the same amount of data as in Condition 1. While the amount of data collected from each person is the same as in the complete data case, the item sample sizes become unbalanced because the Intermediate item group overlaps with the two other item groups and so has twice as much data per item.

In the 75% missing conditions (Conditions 4 through 7), the examinees are again placed into item groups. Once in an item group, however, examinees respond to only a subset of items. Item selection within an item group is always completely random without replacement. In the 75% balanced conditions (Conditions 4 and 5),
a person responds to a random half of their item group, so that the person sample size is maintained at 20 or 40 responses per person. In the 75\% unbalanced conditions (Conditions 6 and 7), examinees respond to 25.0\%, 37.5\%, 50.0\%, 62.5\%, or 75.0\% of their item group. The unbalanced person sample sizes are uniformly distributed and drawn at random, so that on average, there is a 50\% response rate within the item group. Also on average, each person has 20 or 40 data points collected, the same as with the 50\% missingness and the complete data conditions. But in Conditions 6 and 7, the person sample sizes range from 10 to 30 and 20 to 60, respectively.

4.1.2 Models and Items

Data are generated from the 1PL, 2PL, and 3PL IRT models, as well as a condition that includes a subset of nonmonotonic unfolding items. Below are the item parameter generation schemes used for the 1PL, 2PL, and 3PL IRT models.

- $\alpha \sim logN(0.3, 0.46)$
- $\beta_{beg} \sim N(-0.5, 1)$
- $\beta_{adv} \sim N(0.5, 1)$
- $g \sim Beta(5, 17)$

For the 1PL and 2PL, $g$ is set to 0, and additionally for the 1PL, $\alpha$ is set to 1.5 (the mean of its generating distribution). The fourth generating model generates a subset of items from a hyperbolic cosine model (HCM). In this condition, a random 20\% of items are simulated by,

$$\pi_{pi} = \cosh(\alpha_i)/[\cosh(\alpha_i) + \cosh(\theta_p - \beta_i)],$$

which “unfolds,” and the probability of a correct response decreases for higher ability examinees. The HCM condition simulates items that distract high ability examinees.
The α and β parameters are not interpreted as the discrimination and difficulty, although they are drawn from the same distributions. Figure 4.1 shows an example of an HCM item with parameters $\alpha_i = 1.8$ and $\beta_i = 1.0$. A logistic form is clearly incorrect for the HCM item, and, depending on the item parameters, the HCM ICC can be decreasing over a large section of the ability distribution. These items are bad items and are used to evaluate the robustness of estimators to contaminated tests. In the HCM conditions, the remaining 80% of items are simulated from the 2PL model. Even still, a test with 20% of HCM items is extreme in terms of percentage of items that are bad. It would be rare in practice to encounter a well-designed test that contained so many bad items.

4.1.3 Generating Examinees and Item Group Membership

Examinee abilities are simulated from $\mathcal{N}(0,1)$. The Beginning, Intermediate, and Advanced item groups define the missing data. In the complete data condition (Condition 1), the Beginning and Advanced item groups are simulated as described and then combined into a single item bank. Every person responds to all of the items.

In the missing data conditions (Conditions 2 to 7), examinees are placed into one of three item groups. The item group assignment is related to the person’s true ability. Denote the latent item group membership as $\tau_p^*$ and define the process

$$\tau_p^* = \rho \times \theta_p + \sqrt{1 - \rho^2} \times \epsilon_p$$

(4.1)

where $\epsilon_p \sim \mathcal{N}(0,1)$. The manifest item group membership occurs from discretizing $\tau_p^*$ into $\tau_p$, which takes values 1, 2, and 3, representing Beginning, Intermediate, and Advanced item groups. The thresholds $\lambda_1$ and $\lambda_2$ for the discretization are set such that $Pr(\tau^* < \lambda_1) = 1/3$ and $Pr(\tau^* > \lambda_2) = 1/3$, which sets the item groups to have roughly equal membership. The polyserial correlation between the observed item
Figure 4.1. Example of One Item from the Hyperbolic Cosine Model

group membership and latent ability is $\rho$. In the MCAR conditions, $\rho = 0.0$ and the item group membership is completely at random. In the MNAR conditions $\rho = 0.8$, which creates a strong relationship between ability and group membership.

A point of discussion in the IRT literature is how and when missingness due to ability is MAR or MNAR. If the missingness is due to the unobserved $\theta_p$, it is MNAR. However, if the missingness is due to $\hat{\theta}_p$ which is calculated from observed responses to pre-calibrated items, then the missing mechanism is MAR, as it is in CAT. While the generation of $\tau_p^*$ clearly relies on $\theta_p$ and thus is MNAR, it would be simple to change the situation and have the same data generation scheme produce MAR data. If $\tau_p^*$ was, instead, an estimate of student ability based on a pre-test, the missingness caused
by the item groups would be MAR, although the additional measurement error may lower the correlation between true ability and group membership. If item membership was determined by a classroom teacher based on the teacher’s profile of the student that was founded on previous classroom performance, then the missingness would also be MAR. So while the missingness above is labelled MNAR, it could also reasonably arise from a MAR process.

To summarize the data generating conditions, there are seven missing data conditions, two person data sizes (20 or 40), and four generating item response models for a total of 56 data generating conditions. In each condition, the “observed data” is the response data matrix containing generated responses and missing data.

4.2 Study 1: Ranking Examinees

Study 1 compares three ordinal estimators of ability. In all conditions, the baseline that all estimators are trying to attain is the rank of $\theta$. This represents the true rank order of the examinees, which is unknowable in real data. Each set of ordinal ability estimates is evaluated against the baseline by Spearman’s $\rho$, which is the correlation between the rank of $\theta$ and the rank of the ability estimates.

The ordinal ability estimators tested are the ALS-SVD $r$, Proportion Correct $u$, and the 2PL-based MLE, $\hat{\theta}_{2PL}$. Of note, the 2PL ability estimates are estimated from a two-stage estimator that is common in the IRT field. First, the item parameters are estimated with MMLE, then, assuming the item parameters are fixed, the ability estimates are estimated with MLE. In the first stage it is not uncommon to have non-converged results, however, in the simulations there is no significant difference between the resulting rank order of the converged versus non-converged calibrations. Therefore, the below results do not remove these non-converged results. MMLE does not work if an item receives all 0’s or 1’s, so these items are removed. Similarly, MLE cannot estimate the ability of examinees that have perfect responses (all 0’s
or 1’s). When this occurs, those individuals are given an arbitrarily high or low ability estimates, which when taking the ordinal estimate places them at the top or bottom of the ability ranking. Any ties in rank resulting from any estimator are broken randomly. The simulations are run in R with the 2PL item parameters calibrated with the package mirt (Chalmers 2012) and the ALS-SVD computed with the algorithm in Appendix A.

Each condition was replicated 1000 times. Across those 1000 repetitions, the distribution of \( \rho_s \) is summarized in tables below by its mean and its 95% confidence interval calculated by the empirical quantiles of \( \rho_s \) over the 1000 repetitions. To help the interpretability of the \( \rho_s \), Figure 4.2 shows eight bivariate samples of uniformly distributed variables with a sample size of 1000 that vary in their Spearman \( \rho_s \), ranging from 0.5 to 1.0. The results in the tables below range from about 0.50 up to 0.97.

4.2.1 Study 1 Results

Results regarding the ranking of examinees from Conditions 1 to 3 are presented in Tables 4.2-4.4 below with Conditions 4 to 7 appearing in Tables B.1-B.4 in Appendix B. Each table presents a single missing data condition and is organized by generating IRT model, number of items, and estimator. The cells of the table present the mean of \( \rho_s \) and the 95% CI of \( \rho_s \). Higher \( \rho_s \) represents a better estimator and the 95% CI represents the variability of performance over the 1000 repetitions. While \( \rho_s \) has a natural upper bound of 1.0, the 95% CIs are empirical intervals based on the observed quantiles of the results, and thus respect the natural upper bound. The ideal estimator would have \( \rho_s \) near 1.0 and a tight CI. Any source of measurement error lowers \( \rho_s \) and sampling variability from the persons and the items is reflected in the 95% CI. The bottom row of each table presents which estimator was the best overall and the associated percentage indicates the percentage of repetitions in which
that estimator was the best over the 1000 repetitions.

The complete data condition (Condition 1) is presented in Table 4.2. Regardless of estimator, a longer test, which means more data per person, leads to better performance. Also, data generated from the 1PL and 2PL models lead to better estimates than either the 3PL or the HCM models (although recall that the HCM condition has 80% of items generated from the 2PL model).

When data are generated from the 1PL model, proportion correct is the best performing, although differences between it and the other estimators are very small. Under the 1PL model and complete data, proportion correct is a sufficient statistic for ability, so the item level weights computed by ALS-SVD as well as the varying discrimination parameters calibrated by the 2PL IRT model represent random variability in the data. While the 2PL model is over-specified when the data are truly 1PL, over-specification is generally not as large of a concern as under specification.
<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>20</td>
<td>.928 (.912-.940)</td>
<td>.914 (.885-.939)</td>
<td>.853 (.799-.897)</td>
<td>.863 (.796-.909)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.962 (.955-.968)</td>
<td>.955 (.940-.965)</td>
<td>.918 (.894-.939)</td>
<td>.925 (.893-.947)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>20</td>
<td>.921 (.901-.936)</td>
<td>.908 (.873-.935)</td>
<td>.861 (.811-.901)</td>
<td>.850 (.774-.904)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.958 (.950-.965)</td>
<td>.951 (.934-.962)</td>
<td>.923 (.900-.942)</td>
<td>.917 (.879-.943)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>20</td>
<td>.926 (.910-.939)</td>
<td>.926 (.900-.949)</td>
<td>.876 (.827-.913)</td>
<td>.912 (.873-.938)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.961 (.955-.967)</td>
<td>.961 (.949-.971)</td>
<td>.933 (.912-.950)</td>
<td>.954 (.939-.966)</td>
</tr>
<tr>
<td>Best? $u$</td>
<td>20</td>
<td>(99%) $\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (99%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>(96%) $\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td></td>
</tr>
</tbody>
</table>

When the data are generated from the 2PL model, the 2PL IRT model is correctly specified and $\hat{\theta}_{2PL}$ performs better than proportion correct and ALS-SVD. This is also the case with the HCM model, where ALS-SVD and proportion correct struggle with a test contaminated with nonmonotonic items. Interestingly, compared to proportion correct and $\hat{\theta}_{2PL}$, ALS-SVD performs better, although slightly, under the 3PL model than under the HCM model, which is reverse the pattern of the other estimators.

One reason why $\hat{\theta}_{2PL}$ performs well under the HCM model is that the discrimination parameter in the MMLE calibration can be close to zero or even negative for the HCM items. When the estimated discrimination is near zero, the item effectively does not contribute to the ability estimate. When the estimated discrimination is negative, higher ability predicts incorrect responses. This can occur under the HCM.
model, but when it does, the specified 2PL IRT model must still have monotone (in this case, increasing or decreasing) ICCs. In the simulations, the estimated discriminations for HCM models are generally near zero.

For the 2PL IRT model, a near zero discrimination effectively removes that item from the ability estimation. In the case of ALS-SDV, an item can be removed from the ability estimation if the value of $c$ is near zero. However, the ALS-SVD item easiness $c$ is generally not near zero for HCM items. This is understandable, because $c$ is related to $Pr(Y = 1)$ which is clearly above zero for HCM items. So the 2PL model seems better suited to deal with the HCM items.

The 50% MCAR condition (Condition 2) is presented in Table 4.3. Compared to Condition 1, proportion correct’s performance is markedly worse. In comparison, ALS-SVD’s performance is only slightly decreased and $\hat{\theta}_{2PL}$ performs almost identically as it did in Condition 1. Proportion correct’s decreased performance is

### TABLE 4.3

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS CONDITION 2: 50% MCAR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>40</td>
<td>.851 (.753-.918)</td>
<td>.842 (.739-.910)</td>
<td>.792 (.701-.861)</td>
<td>.795 (.678-.873)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.881 (.812-.937)</td>
<td>.876 (.807-.934)</td>
<td>.846 (.773-.903)</td>
<td>.850 (.758-.916)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>40</td>
<td>.910 (.886-.927)</td>
<td>.898 (.869-.925)</td>
<td>.853 (.816-.886)</td>
<td>.844 (.780-.890)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.948 (.935-.959)</td>
<td>.942 (.926-.955)</td>
<td>.916 (.897-.933)</td>
<td>.910 (.880-.932)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>40</td>
<td>.926 (.914-.935)</td>
<td>.924 (.902-.942)</td>
<td>.872 (.837-.902)</td>
<td>.910 (.883-.932)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.961 (.955-.966)</td>
<td>.961 (.952-.968)</td>
<td>.931 (.915-.945)</td>
<td>.954 (.942-.963)</td>
</tr>
<tr>
<td>Best?</td>
<td>40</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
</tr>
</tbody>
</table>
TABLE 4.4

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY
ESTIMATORS CONDITION 3: 50% MNAR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>40</td>
<td>.783 (.576-.903)</td>
<td>.773 (.576-.891)</td>
<td>.704 (.506-.837)</td>
<td>.714 (.507-.859)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.841 (.707-.924)</td>
<td>.829 (.692-.920)</td>
<td>.788 (.666-.884)</td>
<td>.791 (.641-.901)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>40</td>
<td>.912 (.883-.931)</td>
<td>.901 (.861-.929)</td>
<td>.859 (.812-.898)</td>
<td>.847 (.760-.902)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.951 (.937-.961)</td>
<td>.945 (.927-.958)</td>
<td>.922 (.901-.941)</td>
<td>.913 (.869-938)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>40</td>
<td>.876 (.846-.899)</td>
<td>.869 (.813-.914)</td>
<td>.766 (.687-.834)</td>
<td>.839 (.764-.895)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.946 (.935-.954)</td>
<td>.944 (.925-.960)</td>
<td>.887 (.849-.917)</td>
<td>.929 (.902-.951)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Best?</th>
<th>$I$</th>
<th>$r$ (99%)</th>
<th>$r$ (97%)</th>
<th>$r$ (99%)</th>
<th>$r$ (65%)</th>
<th>$\hat{\theta}_{2PL}$ (93%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>r (99%)</td>
<td>r (97%)</td>
<td>r (99%)</td>
<td>r (65%)</td>
<td>$\hat{\theta}_{2PL}$ (93%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>r (89%)</td>
<td>r (55%)</td>
<td>r (100%)</td>
<td>$\hat{\theta}_{2PL}$ (93%)</td>
<td></td>
</tr>
</tbody>
</table>

expected, as proportion correct does not account for item level differences, even if the missingness is completely at random. The patterns of performance of $\hat{\theta}_{2PL}$ and ALS-SVD are similar to that of Condition 1. $\hat{\theta}_{2PL}$ performs better overall and in every repetition, with ALS-SVD being second best.

The 50% MNAR condition (Condition 3) is presented in Table 4.4. Proportion correct performs worst again and worse than it did under MCAR in Condition 2. This result is expected. ALS-SVD performs very similarly to how it performed in Conditions 1 and 2; the missing mechanism does not seem to harm its performance. Notably, $\hat{\theta}_{2PL}$ overall is not as good as ALS-SVD in this condition, except in under the HCM model with $I = 80$. Even when the 2PL model is the generating model, compared to $\hat{\theta}_{2PL}$, ALS-SVD produces a better ranking of the examinees almost all of the time when each examinee responds to only 20 items, and produces a better ranking a little more than half the time when each examinee responds to 40 items.
Under the 3PL model, ALS-SVD is sizably better than $\hat{\theta}_{2PL}$.

For space considerations, Conditions 4 to 7 appear in Appendix B in Tables B.1-B.4. Generally, Condition 4 and 6 (75% MCAR) follow the patterns seen in Condition 2 with more missing data leading to slightly worse performance overall and the balanced person sample size in Condition 4 performing a little better than the unbalanced person sample size in Condition 6. Similarly, Conditions 5 and 7 (75% MNAR) follow the patterns seen in Condition 3. Looking across all seven missing data conditions, Figure 4.3 summarizes the results from Study 1 (except for the HCM generating model). What is most interesting is that even when data are generated from the 1PL and 2PL models, and so the 2PL IRT model is correctly or over-specified, $\hat{\theta}_{2PL}$ can perform worse than ALS-SVD in MNAR conditions. This result motivates Study 1B below.

4.3 Study 1B: At What Point is ALS-SVD Better Than IRT?

Reflecting on the results of Study 1, it is intuitive why proportion correct struggles with missing data and struggles further when that missing data are caused by MNAR. The reason why $\hat{\theta}_{2PL}$ performs well with MCAR data but begins to struggle with MNAR data warrants further explanation and investigation. Recall that the item parameters are calibrated with MMLE. While MMLE, as most likelihood approaches, appropriately handles MCAR and MAR data, the MNAR data of Conditions 3, 5, and 7 violate the ignorability principle (Little and Rubin, 2002).

Because the likelihood specification of the IRT model does not take into account the missing data mechanism, the item parameter estimates can be systematically biased (Rose et al., 2010). The error in item parameter estimates can propagate to the ability estimates in the second stage of the IRT ability estimator. Table 4.5 presents the mean and standard deviation of the bias and root-mean-squared-error (RMSE) of the item parameters of the estimated 2PL IRT model under Conditions
Figure 4.3. Mean of $\rho_s$ Across All Study 1 Missing Data Conditions, 20 Responses per Person

2 and 3 in Study 1. In both conditions, the data are 50% missing, but the missing mechanism changes from MCAR to MNAR. The results below show only the case when the generating model is the 2PL model. So, under Condition 2, the only source
of error is sampling error and under Condition 3, the sources of error are sampling error and missing mechanism misspecification. In both Conditions, the form of the ICC (the 2PL) is correctly specified.

**TABLE 4.5**

**MEAN (AND STANDARD DEVIATION) OF BIAS AND ROOT-MEAN-SQUARED-ERROR OF ITEM PARAMETERS UNDER 50% MCAR AND MNAR**

<table>
<thead>
<tr>
<th>Condition</th>
<th>( I )</th>
<th>Bias of ( \alpha )</th>
<th>RMSE of ( \alpha )</th>
<th>Bias of ( \beta )</th>
<th>RMSE of ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. 50% MCAR</td>
<td>40</td>
<td>0.027 (0.045)</td>
<td>0.083 (0.146)</td>
<td>0.002 (0.039)</td>
<td>0.032 (0.054)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.025 (0.034)</td>
<td>0.056 (0.053)</td>
<td>0.002 (0.035)</td>
<td>0.034 (0.062)</td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td>40</td>
<td>-0.256 (0.047)</td>
<td>0.140 (0.060)</td>
<td>-0.008 (0.102)</td>
<td>0.430 (1.567)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>-0.163 (0.037)</td>
<td>0.084 (0.042)</td>
<td>0.004 (0.060)</td>
<td>0.229 (0.716)</td>
</tr>
</tbody>
</table>

When data are MNAR, the discrimination parameters are, on average, negatively biased. The bias is less with more data per person, but still present. This flatter relationship between ability and the probability of correct response directly increases the measurement error of \( \theta \) (e.g. the test information function). Thus by ignoring non-ignorable missing data, the IRT model, under these specific conditions, compensates by decreasing the measurement precision of the items (by underestimating the item discriminations). This balancing act, where the discrimination parameter absorbs un-modeled error, is a noted phenomenon in IRT models \[\text{Lathrop, 2014}\].
Recall that the item group membership is defined by $\tau$ and is correlated with $\theta$ in the MNAR conditions. When the correlation is 0, MCAR holds and the likelihood specified by MMLE appropriately ignores the missing mechanism. In the MNAR conditions, the polyserial correlation between $\tau$ and $\theta$ is 0.8, and with any non-zero correlation, the MMLE likelihood is incorrect as it does not model the missing data directly (Rose et al., 2010). While a correlation of 0.8 is clearly strong in the context of previous simulation studies, the results of Study 1 do provide, at the minimum, examples of situations where ALS-SVD is a better ordinal estimator of ability than a 2PL IRT model even when the form of the ICC is correctly specified. While the discrete nature of the data generation conditions in Study 1 shows that there is a meaningful difference between a 0.0 and a 0.8 ability/item group correlation, different magnitudes of correlation would lead to different magnitudes of error propagation. Clearly, “the more the correlation differs from zero, the more ignorability is violated” (Holman and Glas, 2005, p.6). It would be expected that a very small, but non-zero correlation, while still incorrectly specifying the missing mechanism, would not lead to meaningful differences. Similarly, as the correlation increases, error that manifests in the ability estimates would also increase.

In short, Study 1 leaves a very important question unanswered. That question is, at what point does the error in the MMLE calibration caused by MNAR propagate enough to the ability estimates to cause concern. As a post-hoc simulation study, Study 1B aims to answer this question by simulating missing data not at discrete correlation values but at all values between 0 and 1.

A second factor that Study 1B considers is the number of items that form the common anchor across item groups. In Study 1 (and Study 2 and 3 below), the Intermediate item group shares half of its items with Beginning item group and half of its items with the Advanced item group. The overlap across item groups is held constant. Study 1B instead varies the amount of overlap for two reasons. First, as
psychometric data are collected from new and complex situations, the sparsity of the response matrix is expected to increase. Second, the amount of overlap is an observable feature of the response matrix. While “overlap” is not defined in this work (although a meaningful index could be developed to include both amount and structure of missingness), it relates to the observed patterns of missing data, which relates to how the common ability scale is formed across the item groups. Study 1B manipulates the amount of overlap in a simple two-group missing data situation to begin to explore its effect. The smaller the overlap between groups, the harder it will be for any ability estimator to set a common scale.

4.3.1 Study 1B Conditions and Results

In Study 1B, $N = 2000$ and examinees respond to 20 or 40 items. There are only two item groups (Beginning and Advanced) each with 1000 examinees, and the item generation is as before for the 1PL, 2PL, and 3PL generating models. Instead of an Intermediate group anchoring the Beginning and Advanced groups, in Study 1B the Beginning and Advanced groups directly share a common set of items. The size of the overlapping set of items is randomly drawn from between 5% and 75% of 20 or 40 items. With 20 responses per person, the overlap ranges from only 1 item to 15 items and with 40 responses per person, ranges from 2 to 30 items. As overlap decreases, the amount of missing data increases. When the two groups overlap at 75%, the resulting response matrix is 20% missing. When the two groups overlap at only 5%, the resulting response matrix is 51% missing. But keep in mind that this two-group, two missing pattern design is in some ways the simplest case of group-level missing data.

Item group membership is generated as before, except that the MNAR correlation is randomly drawn from a uniform distribution from 0.0 to 1.0. To effectively sample from both the overlap and MNAR correlation distributions, Study 1B contains 10000
repetitions. In each, as in Study 1, the rank correlation between the true ability and the estimators ALS-SVD $r$, proportion correct $u$, and the IRT-based $\hat{\theta}_{2PL}$ are recorded.

It is important to realize that by drawing both the overlap and the MNAR correlation at random, comparisons must be made with respect to both factors; it does not make sense to marginalize over one of the factors. In the extreme case that the overlap is 100% (a situation not simulated in Study 1B) the MNAR correlation is irrelevant because all items are shared between the Beginning and Advanced groups, and thus there is no missing data. Because of the dependence of the two factors on each other, results are presented below as the rank correlation $\rho_s$ as a function of one factor conditional on specific values of the other factor.

Figures 4.4 and 4.5 show the smoothed rank correlation as a function of the MNAR correlation at specific ranges of overlap for 20 and 40 responses per person respectively. What is most striking about the results in Figures 4.4 and 4.5 is how ALS-SVD maintains an almost flat line in every panel. This can be interpreted by saying that the accuracy of the ability ranking produced by ALS-SVD does not diminish as a function of the MNAR correlation. The flat relationship exists at all levels of MNAR correlation and at the three overlap slices. Both proportion correct and $\hat{\theta}_{2PL}$ have different levels of performance as a function of the MNAR correlation, and that relationship is different at different levels of overlap. When overlap is between 50% and 75% (which again is confounded with the overall amount of missing data), then the magnitude of the MNAR correlation has a much smaller influence on the performance of proportion correct and $\hat{\theta}_{2PL}$ as when the overlap is between 5% and 25%. The decrease in performance is particularly dramatic with high MNAR correlation and small overlap when the data are generated from the 3PL model. With 40 responses per person, proportion correct and $\hat{\theta}_{2PL}$ are slightly more robust to both high MNAR correlation and low overlap.
Figures 4.6 and 4.7 also show the results of Study 1B, but this time present the mean rank correlation of the ability estimators at discrete levels of overlap, conditional on a specific range of MNAR correlation. In the top row, when the MNAR correlation is less than 0.1 (that is, the missingness is close to ignorable), there is no meaningful effect of overlap on ALS-SVD or $\hat{\theta}_{2PL}$. Proportion correct does suffer with
Figure 4.5. Study 1B: Smoothed $\rho_s$ as a Function of MNAR Correlation at Different Levels of Overlap, 40 Responses Per Person

decreased overlap, which is essentially the same result seen in the MCAR conditions in Study 1. As the level of MNAR correlation increases, especially when it is above 0.7, $\hat{\theta}_{2PL}$ struggles at lower levels of overlap. As before, ALS-SVD shows an almost flat relationship with overlap at all levels of the MNAR correlation.
4.4 Study 2: Nonparametric ICCs

Study 2 focuses on how the choice of ordinal ability estimator used by nonparametric kernel smoothing affects the accuracy of the estimated nonparametric ICCs. Each item’s ICC is estimated nonparametrically using ALS-SVD $r$, rest-ALS-SVD
Figure 4.7. Study 1B: Mean $\rho_s$ as a Function of Overlap at Different Levels of MNAR Correlation, 40 Responses Per Person

$r_{(-i)}$, proportion correct $u$, and rest-proportion-correct $u_{(-i)}$. In addition, the item parameters from the MMLE 2PL calibration are used to estimate a parametric ICC. These five ICC estimates are compared to the true baseline, which is defined by the generating IRT model and item parameters.
The accuracy of the estimated ICCs is evaluated by numerically integrating the squared distance between the true and estimated ICCs. The integration is over the $\mathcal{N}(0, 1)$ $\theta$ distribution and performed at 101 evaluation points placed on the $\mathcal{N}(0, 1)$ density. The square root is then taken. This is the $RISE$ statistic (although for Study 2, $RISE$ is used as a measure of distance between two ICCs, and is not used to test item fit as it is in Study 3). The nonparametric ICCs are evaluated at the 101 $RISE$ integration points. The parametric ICC is calculated at these points as well. Because the integration is over $\mathcal{N}(0, 1)$ and not an equally spaced grid, the $RISE$ statistic places more emphasis on the ICC being accurate in the middle of the $\theta$ distribution, and less emphasis on the tails of the ICC, where the density of $\theta$ is small and differences in ICCs are less meaningful.

The nonparametric ICCs are estimated with a gaussian kernel and a bandwidth parameter equal to $1.06 \times N_i^{-1/5}$, where $N_i$ is the number of examinees that respond to item $i$. For the chosen ordinal ability estimator, the ability estimates are ranked and then placed on the quantiles of $\mathcal{N}(0, 1)$ to mimic the latent scale as explained in Section 1.1.2. The parametric 2PL estimated ICC serves as a “usual practice” baseline to provide a context in which the performance of the nonparametric ICCs can be evaluated. As in Study 1, the 2PL item parameters are estimated with MMLE with the R package mirt (Chalmers, 2012).

Study 2 has 1000 repetitions per condition. For each pair of true and estimated ICCs, the $RISE$ is calculated and stored. Because this is calculated and stored per item, there are between 20 (in Condition 1) to 160 (in Conditions 4-7) $RISE$ values stored per repetition per estimated ICC. The tables below present the average $RISE$ across all items as well as the 95% empirical CI of $RISE$. 
4.4.1 Study 2: Results

Results for Study 2 are presented for Conditions 1 to 3 in Tables 4.6 to 4.8 and for Conditions 4 to 7 in Tables B.5 to B.8 in Appendix B. As values in the tables are in terms of error, smaller values indicated better estimation compared to the true ICC. In Tables 4.6 to 4.8, it is clear that the rest $r_{(-i)}$ and rest $u_{(-i)}$ do not perform as well as their “non-rest” alternatives. Recall too that the rest-estimators are more computationally demanding. While there has been disagreement in the literature on the need for rest scores in theoretical and example-based discussions, this is the first simulation I am aware of that directly compares the accuracy of nonparametric ICCs with rest- and non-rest-scores. It seems that empirically, rest-scores not lead to better results, even in the case of proportion correct under complete data. Interestingly, in Condition 2, rest $u_{(-i)}$ does perform better than $u$, although only in a relative sense as both $u_{(-i)}$ and $u$ produce ICCs with high RISE. This pattern only exists in the MCAR conditions. Because of their general poor performance, Tables B.5 to B.8 present Conditions 4 through 7 without including the rest-estimators. The following text also only discusses the non-rest estimators.

The kernel smoothing computes a weighted average of nearby ability points, so it is expected that some of the difference seen in the rankings in Study 1 will now be averaged out. If the error in the rankings is of a systematic nature, then the nonparametric ICC may be biased. In Condition 1, both nonparametric ICC estimators, ALS-SVD and proportion correct, perform very similarly. Recall that in Study 1, their ability rankings were roughly equal in terms of performance, so it is not a surprise that their performance is so similar here. Compared to the parametric ICC estimated by the 2PL model, the nonparametric ICCs performance is as good when the data are truly from a 3PL model, and worse when data are from the 1PL, 2PL,
TABLE 4.6

MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 1: COMPLETE DATA

<table>
<thead>
<tr>
<th>Estimator</th>
<th>I</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD r</td>
<td>20</td>
<td>.028 (.013-.050)</td>
<td>.032 (.012-.061)</td>
<td>.036 (.014-.064)</td>
<td>.046 (.014-.117)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.027 (.012-.049)</td>
<td>.029 (.012-.054)</td>
<td>.031 (.014-.054)</td>
<td>.038 (.014-.083)</td>
</tr>
<tr>
<td>rest r_{(−i)}</td>
<td>20</td>
<td>.042 (.017-.069)</td>
<td>.046 (.018-.084)</td>
<td>.047 (.020-.085)</td>
<td>.057 (.023-.111)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.035 (.014-.060)</td>
<td>.038 (.016-.067)</td>
<td>.039 (.017-.067)</td>
<td>.045 (.019-.082)</td>
</tr>
<tr>
<td>PropCor u</td>
<td>20</td>
<td>.026 (.011-.046)</td>
<td>.030 (.012-.056)</td>
<td>.037 (.014-.067)</td>
<td>.040 (.013-.100)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.026 (.011-.047)</td>
<td>.029 (.012-.052)</td>
<td>.032 (.014-.055)</td>
<td>.034 (.013-.070)</td>
</tr>
<tr>
<td>rest u_{(−i)}</td>
<td>20</td>
<td>.041 (.016-.067)</td>
<td>.045 (.018-.083)</td>
<td>.049 (.021-.088)</td>
<td>.054 (.021-.103)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.035 (.014-.059)</td>
<td>.037 (.015-.066)</td>
<td>.040 (.018-.069)</td>
<td>.042 (.018-.076)</td>
</tr>
<tr>
<td>2PL</td>
<td>20</td>
<td>.018 (.003-.041)</td>
<td>.019 (.003-.042)</td>
<td>.035 (.008-.084)</td>
<td>.028 (.004-.094)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.018 (.003-.040)</td>
<td>.018 (.003-.042)</td>
<td>.036 (.009-.084)</td>
<td>.028 (.004-.094)</td>
</tr>
</tbody>
</table>

or HCM. In the 3PL condition, the means of RISE are similar across all estimators, but the parametric ICC has a wider 95% CI, indicating increased variability in performance.

When data are 50% MCAR in Condition 2 in Table 4.7, proportion correct has a RISE over twice as high as with complete data. ALS-SVD also shows a larger RISE, but not nearly the same magnitude as proportion correct. The parametric ICC continues to perform well under 50% MCAR, outperforming ALS-SVD regardless of the generating model.

In Condition 3, when data are 50% MNAR, all RISE values are larger than in Condition 2. Reflecting the same pattern seen in Study 1, there are conditions where ALS-SVD produces a better estimated ICC than the 2PL parametric ICC. This even occurs when the data are generated from the 2PL model and examinees respond to
TABLE 4.7

MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 2:

50% MCAR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>I</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD r</td>
<td>40</td>
<td>.045 (.016-.094)</td>
<td>.046 (.016-.098)</td>
<td>.044 (.017-.084)</td>
<td>.057 (.017-.141)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.043 (.015-.089)</td>
<td>.043 (.015-.093)</td>
<td>.040 (.016-.077)</td>
<td>.049 (.016-.108)</td>
</tr>
<tr>
<td>rest r_{(i)}</td>
<td>40</td>
<td>.052 (.018-.106)</td>
<td>.054 (.020-.112)</td>
<td>.053 (.021-.102)</td>
<td>.063 (.024-.125)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.047 (.016-.098)</td>
<td>.048 (.017-.101)</td>
<td>.046 (.018-.088)</td>
<td>.053 (.020-.105)</td>
</tr>
<tr>
<td>PropCor u</td>
<td>40</td>
<td>.090 (.024-.195)</td>
<td>.087 (.024-.200)</td>
<td>.075 (.024-.162)</td>
<td>.086 (.024-.193)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.088 (.026-.179)</td>
<td>.085 (.025-.190)</td>
<td>.072 (.024-.155)</td>
<td>.082 (.025-.183)</td>
</tr>
<tr>
<td>rest u_{(i)}</td>
<td>40</td>
<td>.084 (.027-.173)</td>
<td>.081 (.027-.180)</td>
<td>.069 (.026-.141)</td>
<td>.081 (.030-.172)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.084 (.027-.169)</td>
<td>.082 (.026-.181)</td>
<td>.068 (.025-.144)</td>
<td>.079 (.027-.173)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td>.021 (.004-.051)</td>
<td>.022 (.004-.052)</td>
<td>.038 (.009-.087)</td>
<td>.031 (.004-.095)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.021 (.003-.049)</td>
<td>.021 (.003-.050)</td>
<td>.038 (.010-.087)</td>
<td>.031 (.004-.095)</td>
</tr>
</tbody>
</table>

20 items each. This is a condition in which the 2PL parametric model is correctly specified but the MNAR features of the data bias the parametric ICCs, as explored in Study 1B.

It is important to realize that comparing the ALS-SVD nonparametric ICCs to the 2PL parametric ICC is not a fair comparison because the 2PL parametric ICC uses prior information and constraints that are not available to the nonparametric ICC. And so the major result of Study 2 is that when estimating nonparametric ICCs under missing data (MCAR or MNAR), ALS-SVD is clearly preferred over proportion correct.

Conditions 4 through 7, in Appendix B, follow the same patterns previously established. In all four 75% missing conditions, ALS-SVD is better at estimating
nonparametric ICCs compared to proportion correct. As before, proportion correct performs by far the worse in the 75% missing conditions. Regarding comparing the nonparametric ICCs to the parametric baseline, under MCAR the parametric 2PL ICCs perform the best. Under MNAR, the nonparametric ICCs that use ALS-SVD perform the best. All ICCs are less accurate when the person sample size is unbalanced and when missing data are MNAR.

4.5 Study 3: Item Fit Statistics *RISE*

The nonparametric ICC is then used to test item fit with the bootstrapped procedure based on the *RISE* statistic. Because error in the original examinee rankings (in Study 1) can propagate to the estimation of nonparametric ICCs (in Study 2),
item fit statistics based on those estimated ICCs can also contain error (which is the focus of Study 3). The outcome of interest is the Type I error and power of the RISE item fit statistic based on its bootstrapped p-value. Study 3 uses the same conditions as in Study 1 and Study 2.

In previous work on the RISE fit statistic under complete data, the bootstrap samples are generated from the best fitting (parametric) candidate model and a population distribution for ability. Because there is missing data, and the patterns of missing data are meaningful, the bootstrap procedure needs to be modified. As in the original method, the best fitting candidate model’s item parameters are computed. Then a complete response matrix is generated from those item parameters and from the ability estimates. Finally, if in the original response matrix the person/item interaction does not occur, then the corresponding interaction in the bootstrapped response matrix is removed. This results in a bootstrapped response matrix that has the same pattern of missing data as the original. Notably, this modification requires the use of the ability estimates, instead of simulating new examinees from a population. With missing data there is no easy way around this, because if new examinees are simulated, what missing data pattern should they follow? If the missing data patterns are randomly imposed, or if the missing data patterns are fixed but the abilities are randomly drawn, then all missingness is MCAR. If in the original data matrix the missingness is not MCAR, but the sampling distribution for the fit statistic does follow MCAR, the resulting p-value confounds the item fit test with a test of the ignorability principle of the missing data (Lathrop, 2011).

The bootstrapped p-value is based on 501 bootstrapped samples. The nominal Type I error rate is 0.05. This means that, with 501 bootstrapped samples and a one-tailed test (as the root-squared error is only positive and only extremely large values are concerning), the 95% quantile occurs exactly between the 476 and 477 ordered bootstrap replicates.
There are three candidate IRT models whose parametric form are tested: the 1PL, 2PL, and 3PL models. Because an over specified model is not as large of concern as an under specified model, if the candidate model has more item parameters than the generating model, it is not fit. This results in Table 4.9. Note that in the HCM condition, only 20% of the items are misfitting (recall they are unfolding). Thus, only those misfitting items will be tested for item fit.

TABLE 4.9
GENERATING AND CANDIDATE MODEL TO TEST ITEM FIT

<table>
<thead>
<tr>
<th>Generating Model</th>
<th>Candidate Model</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>Type I error</td>
<td>power</td>
<td>power</td>
<td>power</td>
<td></td>
</tr>
<tr>
<td>2PL</td>
<td>-</td>
<td>Type I error</td>
<td>power</td>
<td>power</td>
<td></td>
</tr>
<tr>
<td>3PL</td>
<td>-</td>
<td>-</td>
<td>Type I error</td>
<td>power</td>
<td></td>
</tr>
</tbody>
</table>

It is important to note that, especially in the case of the 3PL model, simpler IRT models often can explain the observed responses sufficiently well (Han 2012; Wells and Bolt 2008). The main reason for this is that, in the case of the 3PL, the extra parameter that is estimated only affects very low ability examinees. Thus, a simpler model can explain the response curve for the vast majority of examinees and so it would be expected behavior for any fit statistic to not flag the under-specified candidate model. To incorporate this expected behavior, the distance between the
best fitting candidate model and the generating model is measured with a statistic called MISFIT (introduced in Wells & Bolt, 2008):

\[
MISFIT = \sqrt{\sum_{j=1}^{501} w(\theta_j) (\pi_{GenModel,j} - \hat{\pi}_{CandModel,j})^2}
\]  

(4.2)

where \( w(\theta_j) \) is the density of the ability distribution at the \( j \)th point, given by the standard normal density. \( MISFIT \) can then be used to measure how much the best fitting restricted model does not fit the generating model. If \( MISFIT \) is small, power decreases towards 0. Following Wells & Bolt (2008), the tables below use the cutoff of \( MISFIT > 0.020 \) to test power. Note that in the Type I error conditions, \( MISFIT = 0 \) because the candidate model can of course perfectly fit the generated model.

4.5.1 Study 3 Results

The Type I error results for Study 3 are presented in Table 4.10. The values in the table represent the Type I error rate across all items in the specific condition. In Condition 1, \( I = 20 \), this is calculated from 20,000 items (20 items per repetition, 1000 repetitions). In Condition 7 with \( I = 160 \), the rate is calculated from 160,000 items. These large sample sizes result in binomial confidence intervals that generally vary in the third decimal place and so are not reported. This means that if a value in the table is different than 0.05 by even a few thousandths, it is significantly different. For example, in Condition 1, only for \( I = 40 \) and the 2PL and 3PL proportion correct values (0.049 and 0.050) have 95% binomial confidence intervals that cover the nominal value.

First, to connect these results to previous simulation studies on the performance of \( RISE \), in Condition 1, the 2PL Type I error rates for proportion correct are
equal to 0.063 for $I = 20$ and 0.049 for $I = 40$. The Type I error rates reported in [Wells and Bolt (2008)], while using different generating conditions, ranges from 0.04 to 0.06 (the rates are only reported to two decimal places). Even though with complete data the issue of bootstrapping missing data patterns is not relevant, the current implementation of RISE holds the ability estimates fixed while [Wells and Bolt (2008)] simulate new examinees at each bootstrap. Even with this difference in implementation, the overlapping conditions produce agreeable results.

Second, most of the Type I error rates in Table 4.10 are not equal to 0.05. Particularly with 1PL data, the Type I error rates range from too conservative at 0.011, to far too liberal at 0.387. The bold faced elements of the table represent Type I error rates that are not reasonably acceptable under any circumstance. A test with too high Type I error is unusable, although how high is too high depends on the situation and the data analyst. Clearly Type I error rates above 0.2 are too large. However, Type I error rates of 0.05-0.09 may or may not be acceptable depending on the circumstances.

Third, the Type I error rates under the 2PL and 3PL data are more controlled, but vary with the missing data condition. With MCAR missing data, proportion correct and ALS-SVD have Type I error rates that are around 0.04 to 0.05. With MNAR missing data, ALS-SVD often has a too conservative Type I error rate around 0.02 to 0.03 while proportion correct stays around 0.04. While it is clear from Study 1 and Study 2 that ALS-SVD is preferred for ranking examinees and calculating non-parametric ICCs in missing data conditions, those results do not generally translate into better Type I error rates except under the 1PL with MNAR data.

Table 4.11 presents the power rates for testing 1PL and 2PL candidate models against 2PL and 3PL data. Note that the HCM results are reported separately. The elements corresponding to very high Type I error rates (the proportion correct test of
1PL candidate model under MNAR) are crossed out so the test’s high rejection rate is not interpreted as a good thing. In general, ALS-SVD and proportion correct have similar power rates. For the 1PL candidate model (the first two columns), power for ALS-SVD is higher in MCAR conditions, and not comparable with MNAR conditions due to proportion correct’s high Type I error rate. When the candidate model is 2PL and the data are truly 3PL, the power rates are very low for MNAR data and are rarely above 30% with MCAR data. The poor power in missing data conditions is more an effect of missing data on the \textit{RISE} fit statistic and bootstrapping procedure itself, and less an effect of which ordinal estimator is used. This suggests that, except for a test of 1PL fit, the procedure for using \textit{RISE} to test for item fit under missing data needs further development. If a test of the 1PL fit is desired, ALS-SVD should be used due to higher power and reasonably controlled Type I error.

Table 4.12 presents the results for testing the fit of the HCM items for the 1PL, 2PL, and 3PL candidate models. The HCM items are easily rejected as following a 1PL form as power is often above 0.80. For the 2PL and 3PL candidate models, under complete data, the power rates are quite high and range from 0.76 and 0.89. Under missing data, power suffers, particularly under MNAR. For the MCAR conditions power rates drop somewhat, but under MNAR they range from 0.20 to 0.57. Proportion correct generally performs better than ALS-SVD in terms of power to detect HCM items.
# TABLE 4.10

## TYPE I ERROR RATE OF RISE FIT STATISTIC

<table>
<thead>
<tr>
<th>Generating Model/Candidate Model</th>
<th>Condition</th>
<th>1PL/1PL</th>
<th>2PL/2PL</th>
<th>3PL/3PL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>r</td>
<td>u</td>
<td>r</td>
</tr>
<tr>
<td>1. Complete</td>
<td>20</td>
<td>0.032</td>
<td>0.011</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.023</td>
<td>0.013</td>
<td>0.056</td>
</tr>
<tr>
<td>2. 50% MCAR</td>
<td>40</td>
<td>0.075</td>
<td>0.014</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.086</td>
<td>0.017</td>
<td>0.041</td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td>40</td>
<td>0.023</td>
<td>0.244</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.031</td>
<td>0.347</td>
<td>0.025</td>
</tr>
<tr>
<td>4. 75% MCAR Balanced</td>
<td>80</td>
<td>0.076</td>
<td>0.015</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.090</td>
<td>0.018</td>
<td>0.042</td>
</tr>
<tr>
<td>5. 75% MNAR Balanced</td>
<td>80</td>
<td>0.032</td>
<td>0.293</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.035</td>
<td>0.387</td>
<td>0.029</td>
</tr>
<tr>
<td>6. 75% MCAR Unbalanced</td>
<td>80</td>
<td>0.069</td>
<td>0.011</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.085</td>
<td>0.017</td>
<td>0.042</td>
</tr>
<tr>
<td>7. 75% MNAR Unbalanced</td>
<td>80</td>
<td>0.036</td>
<td>0.227</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.036</td>
<td>0.351</td>
<td>0.030</td>
</tr>
</tbody>
</table>
## Table 4.11

**Power of RISE Fit Statistic**

<table>
<thead>
<tr>
<th>Generating Model/Candidate Model</th>
<th>2PL/1PL</th>
<th>3PL/1PL</th>
<th>3PL/2PL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition</strong></td>
<td>$I$</td>
<td>$r$</td>
<td>$u$</td>
</tr>
<tr>
<td>1. Complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>20</td>
<td>0.674</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.712</td>
<td>0.709</td>
</tr>
<tr>
<td>2. 50% MCAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>40</td>
<td>0.511</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.555</td>
<td>0.514</td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>40</td>
<td>0.159</td>
<td>0.937</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.186</td>
<td>0.996</td>
</tr>
<tr>
<td>4. 75% MCAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Balanced</td>
<td>80</td>
<td>0.519</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.562</td>
<td>0.524</td>
</tr>
<tr>
<td>5. 75% MNAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Balanced</td>
<td>80</td>
<td>0.170</td>
<td>0.574</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.189</td>
<td>0.612</td>
</tr>
<tr>
<td>6. 75% MCAR Unbalanced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>80</td>
<td>0.509</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.553</td>
<td>0.506</td>
</tr>
<tr>
<td>7. 75% MNAR Unbalanced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>80</td>
<td>0.165</td>
<td>0.546</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.182</td>
<td>0.606</td>
</tr>
</tbody>
</table>

72
TABLE 4.12

POWER OF RISE FIT STATISTIC FOR HCM ITEMS

<table>
<thead>
<tr>
<th>Generating Model/Candidate Model</th>
<th>Condition</th>
<th>I</th>
<th>r</th>
<th>u</th>
<th>r</th>
<th>u</th>
<th>r</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HCM/1PL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HCM/2PL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HCM/3PL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Complete</td>
<td>20</td>
<td>0.990</td>
<td>0.992</td>
<td>0.763</td>
<td>0.859</td>
<td>0.784</td>
<td>0.874</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.994</td>
<td>0.993</td>
<td>0.849</td>
<td>0.897</td>
<td>0.864</td>
<td>0.908</td>
<td></td>
</tr>
<tr>
<td>2. 50% MCAR</td>
<td>40</td>
<td>0.971</td>
<td>0.965</td>
<td>0.647</td>
<td>0.698</td>
<td>0.662</td>
<td>0.719</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.981</td>
<td>0.978</td>
<td>0.737</td>
<td>0.778</td>
<td>0.757</td>
<td>0.800</td>
<td></td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td>40</td>
<td>0.782</td>
<td>0.974</td>
<td>0.216</td>
<td>0.417</td>
<td>0.180</td>
<td>0.408</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.824</td>
<td>0.985</td>
<td>0.321</td>
<td>0.573</td>
<td>0.284</td>
<td>0.567</td>
<td></td>
</tr>
<tr>
<td>4. 75% MCAR Balanced</td>
<td>80</td>
<td>0.975</td>
<td>0.971</td>
<td>0.612</td>
<td>0.680</td>
<td>0.631</td>
<td>0.706</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.985</td>
<td>0.982</td>
<td>0.731</td>
<td>0.774</td>
<td>0.752</td>
<td>0.798</td>
<td></td>
</tr>
<tr>
<td>5. 75% MNAR Balanced</td>
<td>80</td>
<td>0.814</td>
<td>0.985</td>
<td>0.202</td>
<td>0.402</td>
<td>0.174</td>
<td>0.387</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.834</td>
<td>0.988</td>
<td>0.305</td>
<td>0.558</td>
<td>0.273</td>
<td>0.552</td>
<td></td>
</tr>
<tr>
<td>6. 75% MCAR Unbalanced</td>
<td>80</td>
<td>0.972</td>
<td>0.966</td>
<td>0.625</td>
<td>0.685</td>
<td>0.649</td>
<td>0.714</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.985</td>
<td>0.982</td>
<td>0.725</td>
<td>0.776</td>
<td>0.748</td>
<td>0.802</td>
<td></td>
</tr>
<tr>
<td>7. 75% MNAR Unbalanced</td>
<td>80</td>
<td>0.801</td>
<td>0.979</td>
<td>0.197</td>
<td>0.401</td>
<td>0.171</td>
<td>0.398</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.835</td>
<td>0.986</td>
<td>0.298</td>
<td>0.571</td>
<td>0.264</td>
<td>0.563</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 5

SUMMARY AND CONCLUSIONS

This dissertation provides foundational work in support of using ALS-SVD in psychometric situations. Chapter 3 provides analytical results showing the outcome vectors of ALS-SVD are not only psychometrically desirable, but also psychometrically meaningful. The two resulting vectors $\mathbf{r}$ and $\mathbf{c}$ are shown to be ordinal estimators of examinee ability and of the marginal probability of correct response, respectively.

Following in line with a tradition of analytical work supporting ordinal ability estimation (Douglas, 1997; Stout, 1990; Johnson, 2006), the ability estimator defined by ALS-SVD converges to the true rank order of ability as test length increases. The psychometric value of an ordinal ability estimate is readily apparent. On the item side, the item easiness estimator defined by ALS-SVD also converges to its true value, and the values of $\mathbf{c}$ are shown to be related to the marginal probability of a correct response. Defining $\mathbf{c}$ in psychometric terms is perhaps the most important result of Chapter 3. With this work, there is no good reason ALS-SVD should not be accepted by the psychometric community; it is no longer a foreign data mining tool, but should belong firmly in the psychometrician’s toolbox as the field continues to advance with new and complex data structures.

Chapter 4 provides four related simulation studies that test ALS-SVD against common practice, with particular focus on the detrimental effects of missing data and sparse data sets when the missingness is related to examinee ability. The simulations investigate ranking examinees, estimating nonparametric ICCs with kernel smoothing, and testing item fit with the RISE statistic.
The results of Study 1 show that proportion correct should not be used to rank the examinees if there is missing data, regardless of the missing mechanism at work. Fitting a 2PL IRT model to the items and then estimating the latent ability can appropriately handle MCAR missing data, but struggles with MNAR missing data. Even when the 2PL model is correctly specified, ALS-SVD can perform better in MNAR conditions. ALS-SVD does not perform as well as the parametric ability estimate when the data are truly MCAR, but ALS-SVD appears robust to the MNAR missing mechanism. From a risk/reward stand-point, if a data set does contain missing data, ALS-SVD appears the safer choice because the benefit of the parametric ability estimate when the missing data are truly unrelated to ability is far less than the detriment caused to it by MNAR data.

The results of Study 1B complement and add a finer layer of granularity, particularly in understanding in what situations ALS-SVD can rank examinees better than $\hat{\theta}_{2PL}$, to the Study 1 results. As the amount of missing data increases, and the examinees share fewer and fewer common items, the parametric $\hat{\theta}_{2PL}$ becomes susceptible to error in the examinee rankings as the MNAR correlation increases. The point at which ALS-SVD is a better ordinal estimator is a function of both overlap and MNAR correlation. The most striking result is how the negative effects of higher MNAR correlation and lower overlap on $\hat{\theta}_{2PL}$ are generally not seen with ALS-SVD. Therefore, the more missing data there is in a response matrix, the more aware the analyst must be regarding the missing mechanism when fitting a parametric IRT model.

Truly understanding the cause of missing data becomes futile as data come from more complex situations. When a data set is 75% missing, it seems generally unsafe to blindly assume MCAR. If in any way students or teachers (or movie watchers as in the Netflix challenge) select a subset of items on which to respond, MNAR likely exists. Regarding multidimensional IRT models that model the missing data directly,
as the examinee experience becomes more varied, for example if the assessment is web-based and taken when convenient, compared to all examinees taking the assessment in the same room at the same time, the missingness most likely does not have a homogeneous source, and thus sufficiently modeling it also becomes more difficult.

To use parametric IRT in these new and complex situations seems to require the analyst to know that all possible missing mechanisms, which results in high rates of missing data, are ignorable or correctly modelled. This seems increasingly unlikely with larger and larger data sets. What is known from the results of Study 1B, is that as the amount of missing increases, the effect of ignoring nonignorable missing data is increasingly detrimental to parametric IRT but does not harm the performance of ALS-SVD.

What is quite intriguing is that the patterns and amount of missing data (simply called “overlap” in Study 1B’s minimalist data generation) are observable features of any data set. It will be important for future work to develop a simple index that can quantify not only the amount of missing data but also the size of the common items between examinees that anchor the ability scale. Such an index could be easily calculated from the observed data set. The index could be designed such that as it decreases, the use of parametric IRT becomes more reliant on the assumption of ignorable missing data. If it is clear that the success of parametric IRT for a particular data set requires a strong missing data assumption, that could provide evidence that ALS-SVD might be used instead. Of course, if it is known that the missing data are ignorable, or that the index is not below some threshold, the parametric IRT may be safely used without relying as strongly on assumptions about the missing data.

In Study 2, the examinee ranking produced by ALS-SVD and by proportion correct are used to form nonparametric ICC estimates. When there is any amount of missing data, ALS-SVD produces a more accurate estimate of the true ICC compared to proportion correct. This result is expected because the input to the nonparametric
kernel smoothing is a ranking of the examinees and a better ranking leads to better ICC estimates. One surprising results in Study 2 is that when missing data are MNAR, the nonparametric ICC estimated with ALS-SVD is a better estimate of the true ICC than a correctly specified 2PL IRT model. The parametric model in this situation appears to have many superior qualities in this case (correct ICC form, monotonicity, etc), but under MNAR, the ALS-SVD model is at times better.

The implication is, combining these results with those of Study 1B, that as the amount of missing data increases and is increasingly MNAR, nonparametric estimation of the ICCs using ALS-SVD might be the preferred way to estimate the true ICC. Future work could consider fitting more complicated parametric models, such as the 3PL model, to see if the ALS-SVD nonparametric ICCs are better able to recover educationally meaningful properties of the items, compared to the parametric baseline (for example, recovery of the guessing parameter).

One tangential result regards using rest scores. While the IRT literature is mixed on the use of rest-scores for kernel smoothing, the simulations in Study 2 are the first to directly compare the accuracy of curves with and without rest scores. Even with complete data and only considering the portion correct versus rest-proportion correct rankings, the nonparametric ICCs computed with the rest-scores are more prone to error. Based on these results, and the similar results seen for ALS-SVD and rest-ALS-SVD, if the goal is to minimize error in the estimated nonparametric ICCs, the non-rest-scores should be used.

Study 3 examines the propagation of error from the examinee rankings, to the nonparametric ICCs, and finally to a test of parametric item fit that is based on the nonparametric ICCs. Overall, the error propagation is not as clear as in Study 1 and Study 2 and it seems that the \textit{RISE} fit statistic is influenced by the amount and type of missing data more than anything else. This is because the \textit{RISE} fit statistic implements a bootstrapping procedure, and the bootstrapping must incorporate the
missing data structures that appear in the original data. This is the first study to test the RISE fit statistic under any missing data, and the most meaningful result from Study 3 is that the RISE fit statistic does not appear generally robust to missing data. The results of Study 3 show too liberal and too conservative Type I errors, depending on the missing data amount and condition, and power levels that decrease as the amount of missing data increase. It is important to note though, that the results for the complete data condition replicate those of past research (Wells and Bolt, 2008).

Overall, this work shows that ALS-SVD is a useful and meaningful psychometric tool that has the potential to impact psychometric practices. As the emergence of large data sets continues, the tools used to analyze assessment data will change, and this work provides analytical and empirical evidence that supports ALS-SVD within this context.

Future work could provide advances for the ALS-SVD algorithm; many have already been proposed within the context of data mining. For example, ALS-SVD is a least squares procedure and as such is susceptible to outliers. Even with binary response data, extreme values in $r$ or $c$ could occur if the sample sizes for those elements is very small compared to the average. In this case, it might be wise to impose prior distributions on the output vectors (also known as regularization). Algorithms for regularized ALS exist (Cichocki and Zdunek, 2007), and could be modified for psychometric situations. For example, it is common to impose a $\mathcal{N}(0,1)$ restriction on the ability estimates within IRT. Could a similar regularization in the ALS-SVD improve the ordinal ability estimator? This work would be especially applicable to situations where the sample size varies greatly from person to person.
APPENDIX A

ALS-SVD CODE AND COMPUTATIONAL FEATURES

```r
als <- function(E, max.iter = 100, eps = 1e-5) {
  #E is the response data matrix, coded as 0, 1, and NA

  n <- nrow(E)
p <- ncol(E)
c <- c2 <- matrix(1, p, 1)
r <- r2 <- matrix(apply(E, 1, mean, na.rm=T))
  #initial r values is proportion correct

  n.iter <- 1
  ee<-1

  while(ee > eps & n.iter < max.iter){

    #iteratively fill c vector
    for(i in 1:p){
      mi <- !is.na(E[,i])
r <- r[mi, 1, drop= F]
c2[i,1] <- crossprod(rr, E[mi,i]) / crossprod(rr)
    }

    #iteratively fill r vector
    for(i in 1:n){
      mi <- !is.na(E[i,])
cc <- c2[mi, 1, drop=F]
r2[i,1] <- E[i,mi] %*% cc / crossprod(cc)
    }

    #check convergence
    ee <- max(abs(c-c2), abs(r-r2))
    n.iter <- n.iter + 1
    c <- c2
    r <- r2
  }
  return(list(r, c, n.iter))
}
```

#note: if n.iter == max.iter, convergence was not reached

79
A.1 Computational Features of ALS-SVD

Several points are worth noting regarding the computational features in the algorithm. First, each update of a value in \( \mathbf{c} \) or \( \mathbf{r} \) uses only a small portion of the data. For example, to update \( \mathbf{c}_i \), the algorithm only needs \( \mathbf{y}_i \), or the \( i \)th column of \( \mathbf{Y} \) with missing data removed. The rest of the response matrix is not needed. This presents the opportunity to only store the small slice of data that is currently being used in memory, while the rest of the response matrix does not need to be stored in memory. This would allow the algorithm to be implemented even as the response matrix grows to sizes that cannot be stored directly on the computer’s memory.

Second, while updating an element of \( \mathbf{c} \) or \( \mathbf{r} \) within an iteration, no other values of the \( \mathbf{c} \) or \( \mathbf{r} \) vector are needed. This allows the “for” loops that iteratively update the \( \mathbf{c} \) and \( \mathbf{r} \) vectors to be parallelized across many cores. This would be a relatively easy parallelization because each operation that is sent to a different core is independent of the other operations sent to other cores; their results only need to be combined when they are finished. Due to the separability and independence of the algorithm, ALS-SVD is also a candidate for use with distributed computing systems such as Hadoop.

Finally, the algorithm in above assumes the data are given in a person by item response matrix. This is not the best way to store the data if it contains a large amount of missingness. A more efficient way is in “long form,” where instead of the position in the response matrix being responsible for identifying the person and the item, the person and item identifiers are stored directly. This would entail storing data in three columns labeled “PersonID”, “ItemID”, and “Response.” The major benefit to storing data in long form is that missing data is not stored, and so missing data does not take up space. Of course, the code in Appendix A would need to be modified to accommodate this data storage scheme.
A.2 Timing and Convergence Results

Timing and convergence results from Study 1 appear in Tables A.1 and A.2. First, the definition of convergence is different from ALS-SVD and the MMLE procedure to calibrate 2PL item parameters. For ALS-SVD, convergence occurs when the maximum difference iteration to iteration of any element of the \( r \) and \( c \) vector is less than \( 10^{-5} \) within 100 iterations. For the 2PL item parameters, convergence occurs when the maximum difference of item parameters is less than \( 10^{-4} \) within 500 iterations (the default of mirt). Both use identical machines and only a single core (neither takes advantage of parallel implementations).

For ALS-SVD, the convergence rates are rarely below 100%. For the IRT item calibration, convergence rates decrease with increased missing data and are particularly bad in the HCM data condition. However, the resulting rankings from converged and non-converged repetitions are virtually identical. For example, Condition 6, HCM, \( I = 160 \) has a 53.9% convergence rate for the IRT model. By splitting the ordinal ability estimation results by convergence, the converged 95% CI is .9379 to .9541 and non-converged samples 95% CI is .9382 to .9540. These are very similar, which suggests that while the MMLE calibration did not converge, the results are still useful and are near their optimized values. This similarity is across the other conditions and warrants the combining of converged and non-converged samples as done in the Study 1 results above.

The timing data shows that ALS-SVD is quite a bit faster than the IRT calibration and never averages more than 6 seconds. In many conditions, ALS-SVD converged in .3% to 10% of the time it took the IRT calibration to converge. Even in the largest data condition with the most missing data, the IRT calibration is unlikely to be prohibitively long. It is, however, important to notice the rate at which the IRT
TABLE A.1

COMPUTATIONAL TIME IN SECONDS (AND CONVERGENCE RATES) OF 2PL-IRT AND ALS-SVD FOR 1PL AND 2PL GENERATING DATA

<table>
<thead>
<tr>
<th>Condition</th>
<th>1PL</th>
<th>2PL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( I )</td>
<td>( r )</td>
</tr>
<tr>
<td>1. Complete</td>
<td>20</td>
<td>0.1 (100%)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.2 (100%)</td>
</tr>
<tr>
<td>2. 50% MCAR</td>
<td>40</td>
<td>1.8 (100%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>2.3 (100%)</td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td>40</td>
<td>1.9 (99.5%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>2.3 (100%)</td>
</tr>
<tr>
<td>4. 75% MCAR</td>
<td>80</td>
<td>4.1 (100%)</td>
</tr>
<tr>
<td>Balanced</td>
<td>160</td>
<td>5.3 (100%)</td>
</tr>
<tr>
<td>5. 75% MNAR</td>
<td>80</td>
<td>4.5 (100%)</td>
</tr>
<tr>
<td>Balanced</td>
<td>160</td>
<td>5.9 (100%)</td>
</tr>
<tr>
<td>6. 75% MCAR</td>
<td>80</td>
<td>4.1 (100%)</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>160</td>
<td>5.5 (100%)</td>
</tr>
<tr>
<td>7. 75% MNAR</td>
<td>80</td>
<td>4.4 (100%)</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>160</td>
<td>5.8 (100%)</td>
</tr>
</tbody>
</table>

calibration increases in computational time. As previously discussed, the ALS-SVD algorithm has features that naturally scale as the data increase in size, and that is evidenced by these results.

With even larger data sets (tens or hundreds of thousands of examinees and
TABLE A.2

COMPUTATIONAL TIME IN SECONDS (AND CONVERGENCE RATES) OF 2PL-IRT AND ALS-SVD FOR 3PL AND HCM GENERATING DATA

<table>
<thead>
<tr>
<th>Condition</th>
<th>$I$</th>
<th>$r$</th>
<th>$\hat{\theta}_{2PL}$</th>
<th>$r$</th>
<th>$\hat{\theta}_{2PL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Complete</td>
<td>20</td>
<td>0.1 (100%)</td>
<td>3.5 (99.5%)</td>
<td>0.1 (100%)</td>
<td>4.9 (96.6%)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.1 (100%)</td>
<td>9.2 (98.1%)</td>
<td>0.2 (100%)</td>
<td>12.3 (95.1%)</td>
</tr>
<tr>
<td>2. 50% MCAR</td>
<td>40</td>
<td>1.7 (100%)</td>
<td>12.0 (95.7%)</td>
<td>1.8 (99.9%)</td>
<td>18.4 (91.6%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>2.1 (100%)</td>
<td>40.6 (93.7%)</td>
<td>2.2 (100%)</td>
<td>64.2 (80.2%)</td>
</tr>
<tr>
<td>3. 50% MNAR</td>
<td>40</td>
<td>1.6 (100%)</td>
<td>10.5 (94.5%)</td>
<td>1.7 (99.9%)</td>
<td>14.6 (91.1%)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>2.0 (100%)</td>
<td>30.8 (95.1%)</td>
<td>2.1 (100%)</td>
<td>40.1 (96.5%)</td>
</tr>
<tr>
<td>4. 75% MCAR</td>
<td>80</td>
<td>3.6 (100%)</td>
<td>28.9 (92.4%)</td>
<td>3.8 (100%)</td>
<td>56.7 (70.8%)</td>
</tr>
<tr>
<td>Balanced</td>
<td>160</td>
<td>5.0 (100%)</td>
<td>109.9 (84.3%)</td>
<td>5.2 (100%)</td>
<td>173.9 (54.3%)</td>
</tr>
<tr>
<td>5. 75% MNAR</td>
<td>80</td>
<td>3.7 (100%)</td>
<td>28.8 (90.1%)</td>
<td>4.0 (100%)</td>
<td>42.9 (80.5%)</td>
</tr>
<tr>
<td>Balanced</td>
<td>160</td>
<td>5.1 (100%)</td>
<td>81.0 (90.8%)</td>
<td>5.4 (100%)</td>
<td>110.7 (91.4%)</td>
</tr>
<tr>
<td>6. 75% MCAR</td>
<td>80</td>
<td>3.7 (100%)</td>
<td>28.5 (95.5%)</td>
<td>3.9 (100%)</td>
<td>57.1 (73.8%)</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>160</td>
<td>5.0 (100%)</td>
<td>112.6 (84.7%)</td>
<td>5.3 (100%)</td>
<td>179.0 (53.9%)</td>
</tr>
<tr>
<td>7. 75% MNAR</td>
<td>80</td>
<td>3.8 (100%)</td>
<td>29.2 (88.7%)</td>
<td>4.0 (100%)</td>
<td>39.4 (85.4%)</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>160</td>
<td>5.2 (100%)</td>
<td>81.3 (91.0%)</td>
<td>5.4 (100%)</td>
<td>111.2 (91.1%)</td>
</tr>
</tbody>
</table>

thousands of items), the computation time between IRT calibration and ALS-SVD can be hours compared to minutes. For example, in a very small follow up simulation, under Condition 1, when $N = 12,000$ and $I = 100$, ALS-SVD took under 2.8 seconds and the 2PL IRT model took 182.3 seconds. When $N = 90,000$ and $I = 1,000$,
ALS-SVD took 122.7 seconds and the 2PL IRT model took 64,663.7 seconds (almost 18 hours). Neither of the programs are designed specifically for big data, so the timing comparison is not exactly fair in that it does not compare the minimum computational operations of the most efficiently implemented algorithms. Still, the results are striking.
APPENDIX B

ADDITIONAL TABLES
## TABLE B.1

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 4: 75% MCAR BALANCED

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>80</td>
<td>.846 (.786-.897)</td>
<td>.834 (.763-.886)</td>
<td>.788 (.723-.842)</td>
<td>.790 (.706-.849)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.879 (.832-.919)</td>
<td>.873 (.826-.915)</td>
<td>.844 (.793-.886)</td>
<td>.845 (.788-.895)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>80</td>
<td>.910 (.894-.923)</td>
<td>.897 (.875-.916)</td>
<td>.853 (.828-.877)</td>
<td>.845 (.800-.877)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.948 (.938-.955)</td>
<td>.941 (.931-.951)</td>
<td>.916 (.902-.928)</td>
<td>.909 (.888-.926)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>80</td>
<td>.926 (.917-.934)</td>
<td>.923 (.909-.937)</td>
<td>.873 (.850-.893)</td>
<td>.911 (.892-.928)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.961 (.958-.965)</td>
<td>.961 (.955-.966)</td>
<td>.932 (.920-.942)</td>
<td>.953 (.946-.961)</td>
</tr>
<tr>
<td>Best?</td>
<td>80</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (100%)</td>
</tr>
</tbody>
</table>

## TABLE B.2

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 5: 75% MNAR BALANCED

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>80</td>
<td>.781 (.650-.867)</td>
<td>.764 (.641-.858)</td>
<td>.701 (.574-.808)</td>
<td>.705 (.536-.815)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.836 (.746-.903)</td>
<td>.827 (.741-.896)</td>
<td>.784 (.693-.855)</td>
<td>.786 (.674-.863)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>80</td>
<td>.912 (.894-.926)</td>
<td>.899 (.875-.921)</td>
<td>.861 (.829-.889)</td>
<td>.849 (.791-.888)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.951 (.942-.958)</td>
<td>.944 (.932-.954)</td>
<td>.923 (.908-.935)</td>
<td>.915 (.890-.934)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>80</td>
<td>.877 (.858-.892)</td>
<td>.870 (.833-.902)</td>
<td>.770 (.713-.817)</td>
<td>.842 (.796-.883)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.946 (.939-.952)</td>
<td>.945 (.930-.956)</td>
<td>.888 (.864-.909)</td>
<td>.931 (.911-.946)</td>
</tr>
<tr>
<td>Best?</td>
<td>80</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (67%)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>$r$ (94%)</td>
<td>$\hat{\theta}_{2PL}$ (53%)</td>
<td>$r$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (98%)</td>
</tr>
</tbody>
</table>
### TABLE B.3

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 6: 75% MCAR UNBALANCED

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>80</td>
<td>.835 (.769-.884)</td>
<td>.824 (.764-.875)</td>
<td>.773 (.713-.826)</td>
<td>.772 (.693-.833)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.873 (.825-.914)</td>
<td>.865 (.814-.908)</td>
<td>.835 (.785-.875)</td>
<td>.835 (.779-.882)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>80</td>
<td>.899 (.882-.913)</td>
<td>.885 (.863-.905)</td>
<td>.837 (.808-.862)</td>
<td>.827 (.780-.862)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.941 (.932-.950)</td>
<td>.934 (.922-.945)</td>
<td>.906 (.892-.919)</td>
<td>.899 (.875-.917)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>80</td>
<td>.916 (.907-.925)</td>
<td>.912 (.895-.927)</td>
<td>.857 (.829-.881)</td>
<td>.897 (.874-.914)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.956 (.951-.960)</td>
<td>.955 (.947-.961)</td>
<td>.922 (.910-.933)</td>
<td>.946 (.938-.954)</td>
</tr>
</tbody>
</table>

**Best?**

<table>
<thead>
<tr>
<th>$I$</th>
<th>$\hat{\theta}_{2PL}$ (100%)</th>
<th>$\hat{\theta}_{2PL}$ (100%)</th>
<th>$\hat{\theta}_{2PL}$ (100%)</th>
<th>$\hat{\theta}_{2PL}$ (100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
</tr>
<tr>
<td>160</td>
<td>$r$ (92%)</td>
<td>$\hat{\theta}_{2PL}$ (57%)</td>
<td>$r$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (99%)</td>
</tr>
</tbody>
</table>

### TABLE B.4

MEAN OF SPEARMAN $\rho_s$ (AND 95% CI) OF ORDINAL ABILITY ESTIMATORS; STUDY 1 CONDITION 7: 75% MNAR UNBALANCED

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropCor $u$</td>
<td>80</td>
<td>.766 (.637-.858)</td>
<td>.748 (.625-.845)</td>
<td>.680 (.542-.780)</td>
<td>.685 (.539-.803)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.826 (.736-.895)</td>
<td>.816 (.725-.882)</td>
<td>.767 (.677-.841)</td>
<td>.768 (.656-.852)</td>
</tr>
<tr>
<td>ALS-SVD $r$</td>
<td>80</td>
<td>.901 (.880-.916)</td>
<td>.888 (.860-.909)</td>
<td>.845 (.810-.874)</td>
<td>.833 (.775-.874)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.944 (.934-.952)</td>
<td>.937 (.924-.948)</td>
<td>.913 (.896-.926)</td>
<td>.904 (.878-.924)</td>
</tr>
<tr>
<td>$\hat{\theta}_{2PL}$</td>
<td>80</td>
<td>.865 (.844-.882)</td>
<td>.859 (.822-.890)</td>
<td>.754 (.700-.803)</td>
<td>.828 (.780-.871)</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>.939 (.932-.947)</td>
<td>.938 (.925-.950)</td>
<td>.878 (.851-.901)</td>
<td>.923 (.904-.939)</td>
</tr>
</tbody>
</table>

**Best?**

<table>
<thead>
<tr>
<th>$I$</th>
<th>$r$ (100%)</th>
<th>$r$ (100%)</th>
<th>$r$ (100%)</th>
<th>$r$ (63%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (100%)</td>
<td>$r$ (63%)</td>
</tr>
<tr>
<td>160</td>
<td>$r$ (92%)</td>
<td>$\hat{\theta}_{2PL}$ (57%)</td>
<td>$r$ (100%)</td>
<td>$\hat{\theta}_{2PL}$ (99%)</td>
</tr>
</tbody>
</table>
### TABLE B.5

**MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 4: 75% MCAR BALANCED**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD $r$</td>
<td>40</td>
<td>.043 (.016-.088)</td>
<td>.045 (.016-.093)</td>
<td>.043 (.017-.081)</td>
<td>.055 (.017-.136)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.042 (.015-.086)</td>
<td>.043 (.015-.090)</td>
<td>.039 (.016-.077)</td>
<td>.048 (.016-.105)</td>
</tr>
<tr>
<td>PropCor $u$</td>
<td>40</td>
<td>.086 (.026-.174)</td>
<td>.085 (.027-.182)</td>
<td>.073 (.025-.150)</td>
<td>.082 (.026-.171)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.086 (.028-.167)</td>
<td>.084 (.027-.179)</td>
<td>.071 (.025-.147)</td>
<td>.080 (.027-.170)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td>.021 (.003-.050)</td>
<td>.022 (.004-.051)</td>
<td>.038 (.009-.086)</td>
<td>.031 (.004-.095)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.020 (.003-.048)</td>
<td>.021 (.003-.049)</td>
<td>.038 (.010-.086)</td>
<td>.030 (.004-.095)</td>
</tr>
</tbody>
</table>

### TABLE B.6

**MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 5: 75% MNAR BALANCED**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$I$</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD $r$</td>
<td>40</td>
<td>.054 (.016-.157)</td>
<td>.060 (.015-.179)</td>
<td>.056 (.016-.144)</td>
<td>.082 (.018-.252)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.054 (.013-.161)</td>
<td>.058 (.014-.174)</td>
<td>.055 (.014-.148)</td>
<td>.073 (.016-.214)</td>
</tr>
<tr>
<td>PropCor $u$</td>
<td>40</td>
<td>.105 (.022-.226)</td>
<td>.103 (.023-.263)</td>
<td>.086 (.025-.211)</td>
<td>.113 (.027-.267)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.106 (.021-.223)</td>
<td>.103 (.022-.257)</td>
<td>.085 (.023-.207)</td>
<td>.107 (.025-.257)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td>.078 (.011-.185)</td>
<td>.078 (.010-.209)</td>
<td>.088 (.017-.214)</td>
<td>.091 (.013-.218)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.049 (.006-.138)</td>
<td>.049 (.006-.147)</td>
<td>.067 (.012-.174)</td>
<td>.065 (.007-.176)</td>
</tr>
</tbody>
</table>
### TABLE B.7

**MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 6:**

**75% MCAR UNBALANCED**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>I</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD r</td>
<td>40</td>
<td>.043 (.016-.087)</td>
<td>.045 (.016-.091)</td>
<td>.043 (.017-.081)</td>
<td>.055 (.017-.139)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.042 (.015-.085)</td>
<td>.043 (.015-.088)</td>
<td>.039 (.016-.075)</td>
<td>.048 (.016-.105)</td>
</tr>
<tr>
<td>PropCor u</td>
<td>40</td>
<td>.087 (.027-.176)</td>
<td>.085 (.027-.182)</td>
<td>.074 (.026-.150)</td>
<td>.084 (.026-.174)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.086 (.028-.167)</td>
<td>.084 (.028-.179)</td>
<td>.071 (.025-.146)</td>
<td>.080 (.027-.170)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td>.021 (.003-.050)</td>
<td>.021 (.003-.051)</td>
<td>.038 (.009-.086)</td>
<td>.031 (.004-.095)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.020 (.003-.048)</td>
<td>.021 (.003-.049)</td>
<td>.038 (.010-.086)</td>
<td>.030 (.004-.094)</td>
</tr>
</tbody>
</table>

### TABLE B.8

**MEAN (AND 95% CI) OF RISE OF ESTIMATED ICCS CONDITION 7:**

**75% MNAR UNBALANCED**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>I</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>HCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS-SVD r</td>
<td>40</td>
<td>.055 (.015-.164)</td>
<td>.061 (.016-.182)</td>
<td>.057 (.016-.146)</td>
<td>.083 (.019-.260)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.054 (.013-.164)</td>
<td>.058 (.014-.174)</td>
<td>.055 (.014-.148)</td>
<td>.074 (.016-.221)</td>
</tr>
<tr>
<td>PropCor u</td>
<td>40</td>
<td>.105 (.023-.224)</td>
<td>.102 (.024-.261)</td>
<td>.087 (.026-.209)</td>
<td>.112 (.028-.263)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.105 (.021-.222)</td>
<td>.102 (.022-.255)</td>
<td>.085 (.024-.208)</td>
<td>.107 (.026-.257)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td>.077 (.011-.184)</td>
<td>.076 (.010-.205)</td>
<td>.086 (.017-.211)</td>
<td>.090 (.013-.216)</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.048 (.006-.137)</td>
<td>.048 (.006-.145)</td>
<td>.066 (.012-.173)</td>
<td>.064 (.007-.175)</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


This document was prepared & typeset with pdflatex, and formatted with ndiss2e classfile (v3.2013[2013/04/16]) provided by Sameer Vijay and updated by Megan Patnott.