ROBUST LATENT ABILITY ESTIMATION BASED ON ITEM RESPONSE INFORMATION AND MODEL FIT

by

Hotaka Maeda

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin-Milwaukee

August 2017

ABSTRACT

ROBUST LATENT ABILITY ESTIMATION BASED ON ITEM RESPONSE INFORMATION AND MODEL FIT

by

Hotaka Maeda

The University of Wisconsin-Milwaukee, 2017
Under the Supervision of Professor Bo Zhang

Aberrant testing behaviors may result in inaccurate person trait estimation. To counter its effects, a new robust ability estimation procedure called downweighting of aberrant responses estimation (DARE) is developed. This procedure downweights both uninformative items and model-misfitting response patterns. The purpose of this study is to present DARE and to evaluate its performance against other robust methods, including biweight (Mislevy & Bock, 1982) and biweight-MAP (BMAP; Maeda & Zhang, 2017b). The traditional maximum likelihood (MLE) and maximum a-posteriori (MAP) methods are also included as baseline methods. A Monte Carlo simulation is conducted with the design variables being test length, type of aberrant behaviors, percentage of aberrant examinees, and percentage of aberrant items. Person-fit analyses using l_z^* (Snijders, 2001) and H^T (Sijtsma, 1986) are incorporated as a realistic initial step to determine the aberrant examinees that might benefit from robust estimation methods. Results showed that DARE effectively decreased the root-mean-squared-error (RMSE) and bias of the estimates compared to MAP among examinees detected using the l_z^* at the .01 α cutoff. DARE was the most accurate method in many conditions involving aberrant behavior when the test length was 40 or 60 items. At 20 items, all robust methods were ineffective. DARE performs well when 1) a high-achieving examinee show a mild spuriously low scoring behavior, or 2) a

low-achieving examinee show a mild spuriously high scoring behavior. When used appropriately, DARE is superior to all pre-existing methods in limiting the negative consequences of aberrant behavior.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to thank everyone who supported me throughout my entire academic life; my friends, family, peers, co-workers, advisors, supervisors, instructors, professors, and everyone on my CV. There are too many people to list and thank, but I will name one person. Bruya, thank you so much for exposing me to the world of research and tirelessly encouraging me to pursue a graduate degree and a PhD. Without you, I may have never realized that academic endeavors can be extremely exciting and enjoyable, that I really like research and statistics, and that obtaining a PhD was even an option. Thank you for the life-changing inspiration!

CHAPTER 1: INTRODUCTION

Problem

In many educational and psychological measurement situations, accurately estimating an examinee's level on the latent trait of interest is of utmost importance. Item response theory (IRT) aims to achieve such a goal by modeling the interaction between an examinee and a test item with a probabilistic function. Like all model-based approaches, the accuracy of IRT trait estimation depends on the degree of model-data fit. For example, for examinees who have cheated on some items, a standard IRT model would not be able to account for such aberrant response patterns, resulting in an overestimate of the person trait level. As another example, younger students tend to make careless mistakes, such as inadvertently misreading the instruction for an item. Using a standard model for them would lead to an underestimate of the latent trait. Aberrant test behaviors contaminate item responses and make the otherwise routine task of person trait estimation quite challenging for both aberrant and non-aberrant examinees.

Aberrant test behaviors can manifest in numerous ways, such as cheating, guessing, fatigue, carelessness, excessive creativeness, misunderstanding the instructions, test anxiety, high/low motivation, clinical pathology, tendency to select extreme options, working too methodically, or ignoring negatively worded items (Meijer & Sijtsma, 2001; Rupp, 2013). Some of these behaviors may result in test scores that are too high, such as cheating and lucky guessing, while others such as fatigue or excessive test anxiety may result in spuriously low test scores. Some behaviors may affect the entire test (e.g., clinical pathology) while others may affect only a portion of the test (e.g., fatigue).

Faced with the possibility of aberrant behaviors, one approach is to identify the examinees with aberrant responses and remove them from subsequent analyses. In that regard,

person-fit analyses come into play. Numerous person-fit statistics have been proposed to evaluate whether a response pattern fits a test model (Meijer & Sijtsma, 2001). These statistics provide a statement about the appropriateness of a measurement model. In the event of model-data misfit, however, most person-fit statistics are unable to reveal the nature of the aberrant behaviors.

One exception is the analyses based on the person response function (PRF; Trabin & Weiss, 1983). The PRF gives the probability of a correct answer for an individual with a fixed ability as a function of item difficulty. For a given ability level, as item difficulty increases, the probability of a correct response should decrease. Therefore, the PRF should be non-increasing when the model fits the data. By comparing the observed and expected PRF, one may be able to determine the general pattern of person misfit. For example, misfit of difficult items may provide evidence that the examinee has obtained correct answers through cheating or lucky guessing. On the contrary, misfit of easy items may indicate careless errors. Statistical procedures have been developed to test the non-increasingness of the PRF (e.g., Emons, Sijtsma, & Meijer, 2005; Sijtsma & Meijer, 2001), and to quantify its slope where a steep decreasing angle indicates good person-fit (e.g., Reise, 2000; Strandmark & Linn, 1987). The combination of 1) global person-fit testing, 2) graphical PRF examination, and then 3) local PRF examinations has been proposed as the most comprehensive person-fit analysis (Emons et al., 2005).

Once aberrant response patterns have been detected, the irregularities in the contaminated data can be modeled using robust ability estimation methods, which downweight items that contain little information (Mislevy & Bock, 1982; Schuster & Yuan, 2011; Waller, 1974). These methods can be used to reduce the influence of potential aberrant responses while still retaining most of the trait information embedded in the responses. Recently, Maeda and Zhang (2017b)

proposed Bayesian alternatives to the existing robust estimation methods that downweight items with little information. They showed in their simulation study that the Bayesian extension often substantially improves the estimation accuracy.

Purpose

Existing robust estimation methods have been shown to be useful in reducing the ability estimation bias due to aberrant responding, but their effectiveness is limited and measurement error remains high for many testing conditions (Maeda & Zhang, 2017b; Meijer & Nering, 1997; Schuster & Yuan, 2011). A particular limitation of current methods is that they downweight all uninformative response items, which may cause high loss of information among the non-aberrant responses. Meanwhile, the PRF literature provides potential techniques for identifying aberrant responses. By drawing insights from both the robust estimation and PRF literature, it may be possible to create a procedure that downweights uninformative items as well as items with misfitting observed responses. Those responses that do not fit the model may be particularly likely to be aberrant. Hence, the main purpose of this study is to develop such a procedure and to evaluate its effectiveness in handling various aberrant testing results.

Significance

The importance of this research is fourfold. First, ability estimates are used in high stake decisions like proficiency classification and group comparison. This study explores new ways to provide more accurate ability estimates for examinees with aberrant behaviors when re-testing or dropping the score is unfeasible. Any improvement in ability estimation can significantly improve the utility of test scores. Second, when ability estimates are contaminated by aberrance, they tend to lower the effectiveness of parametric person-fit statistics that rely on accurate ability estimates (e.g., Meijer & Nering, 1997; Reise, 1995). Improving ability estimation may improve

the identification of aberrant behaviors. In a recent example (Maeda & Zhang, 2017a), the power of the copying index omega (Wollack, 1997) increased with better ability estimation. Third, the detection and removal of aberrant examinees may improve item parameter estimation, which in turn will improve ability estimation for non-aberrant examinees, who are usually the overwhelming majority of test takers. Finally, the proposed method may help identify the specific item responses that are aberrant. This information can be of interest in other analyses, such as exploring the source of aberrance.

CHAPTER 2: LITERATURE REVIEW

Studying Aberrant Testing Behaviors

To systematically study the effects of aberrant behaviors on ability estimation, the process in which an aberrant behavior may occur needs to be operationalized. Only then can Monte Carlo simulations be designed to emulate aberrant behaviors in real testing conditions. So far, researchers have come up with a wide variety of operationalized definitions of aberrant behaviors. According to Rupp (2013), the simulation of aberrant responses in previous studies often aim to answer the following questions:

1. How many or what percentage of persons respond aberrantly?
2. What kinds of persons respond aberrantly?
3. To how many items do they respond aberrantly?
4. To what kinds of items do they respond aberrantly?
5. How do they respond aberrantly to selected items?

In theory, the higher the percentage of aberrant examinees, the larger the error in item parameter and ability estimation and the lower the power in detecting aberrant response patterns. This has been confirmed by Karabastos (2003) where the percent of aberrant respondents were simulated as 5%, 10%, 25%, and 50%. Both sensitivity and specificity in detecting aberrant response patterns decreased as more examinees were contaminated. While 1% (e.g., Armstrong & Shi, 2009) to 100% (e.g., de la Torre & Deng, 2008) aberrant examinees has been studied, the most typical cases seem to be around 10% (Rupp, 2013).

In addition, previous research has examined the consequences of aberrant behavior across all ability levels (e.g., Cui & Leighton, 2009; Glas & Dagohoy, 2007) as well as in specific ranges (e.g., Meijer, 1996; Zhang & Walker, 2008). For instance, aberrant behaviors that tend to

cause spuriously high scores (e.g., cheating and lucky guessing) are more likely to occur with examinees with low ability, though medium or high ability examinees may also make lucky guesses or even cheat. In that sense, examining results by ability levels may be particularly valuable for revealing the nature of the aberrant behaviors (e.g., de la Torre & Deng, 2008; Drasgow, Levine, & McLaughlin, 1987; Meijer & Nering, 1997).

Once target examinees have been selected, aberrant behavior must be operationalized. The severity of aberrant behavior has high impact on both the power of detecting aberrant examinees and the accuracy of ability estimation (Rupp, 2013), thus is an important design variable in simulation studies. In practice, aberrance is often simulated by imposing a certain value of conditional probability of a correct response. While some researchers alter all test items (e.g., Schuster & Yuan, 2011), others select some items either randomly (e.g., Drasgow et al., 1987; Levine & Rubin, 1979) or deterministically (e.g., Karabastos, 2003). Researchers usually study only one type of aberrant behavior for one examinee (Rupp, 2013). Although multiple aberrant behaviors (e.g., cheating plus misreading instructions) can occur simultaneously in practice, these kinds of behaviors are difficult to simulate.

For example, Karabastos (2003) simulated cheating by imputing 18% of the most difficult items as correct by setting $P_i^* = 1$ where P_i^* is the probability of a correct response for an aberrant item i. In contrast, careless errors were operationalized by assigning $P_i^* = .5$ to 41% of the easiest items. In these cases, aberrant behavior was modeled as dependent on item difficulty. Alternatively, Schuster and Yuan (2011) used Copas's (1988) model to simulate aberrance for all items:

$$P_i^* = (1 - \gamma)P_i + \gamma Q_i, \quad (1)$$

where, P_i represents the probability of a correct response based on a measurement model, $Q_i = 1 - P_i$, and γ is a value from 0 to 1 that represents the severity of aberrant behavior. Liu, Douglas, and Henson (2009) used a similar approach except Q_i was replaced with A which was fixed as either 1 or 0 to simulate spuriously high or low scores, respectively. It is worth noting that this method of simulating aberrant responses does not necessarily change the original model-generated response.

One interesting observation is that all the above aberrance simulation methods have a differential effect as a function of item difficulty. This is true regardless of whether aberrant items are selected by item difficulty or not. For example, using $\gamma = .2$ in Copas's (1988) model, if $P_i = .1$, then $P_i^* = .26$. Comparatively, if $P_i = .5$, then $P_i^* = .5$. Therefore, $P_i^* > P_i$ for very difficult items (e.g., $P_i = .1$), while $P_i^* = P_i$ for items with medium difficulty (e.g, $P_i = .5$). As another example, if P_i^* is fixed at .2, easier items (e.g., $P_i = .9$) will be affected much more severely than harder items (e.g., $P_i = .1$), as $|.9 - .2| > |.1 - .2|$.

For this reason, aberrant behavior may cause item misfit as a function of item difficulty in both simulation studies and real testing situations. In fact, person response function (PRF; Trabin & Weiss, 1983) is grounded on this idea. Examining the model fit of the items by their difficulty could hint at the nature of the aberrant behavior that may have caused the misfitting response pattern. For example, misfit of difficult items may provide partial evidence that the examinee has obtained the correct answers through cheating or lucky guessing. Evaluating model fit using the PRF will be further discussed in the next section.

Person-Fit Statistics

Person-fit, also known as person appropriateness, refers to the fit of a response pattern to a test model (Meijer & Sijtsma, 2001). Person-fit methods aim to identify model misfit at the individual level, which is distinctly different from the overall fit of a model or the fit of an item to all examinees. That is to say, overall model fit does not require all persons to fit. Person-fit methods strive to identify examinees who have responded aberrantly by identifying atypical response patterns. Importantly, a person detected as misfitting according to a person-fit statistic is not necessarily aberrant. This is because non-aberrant examinees can be erroneously detected as misfitting, while some aberrant examinees can remain undetected. In this sense, the term "misfitting person" is distinct from "aberrant person". To detect the myriad of possible aberrant behavior patterns, a large number of person fit statistics have been proposed and studied for various measurement models. These statistics can be either parametric or non-parametric.

Parametric Person-Fit Statistics

Parametric person-fit statistics detect aberrant examinees by using estimates from the IRT model. The most general unidimensional IRT model is the 3-parameter logistic (3PL) model (Birnbaum, 1968), expressed as

$$P(X_i = 1|\theta) = c_i + \frac{1-c_i}{1+\exp[-1.7a_i(\theta-b_i)]}, \quad (2)$$

where $X_i = 1$ indicates a correct response to item i, θ is the ability level, a_i is the discrimination parameter, b_i is the difficulty parameter, c_i is the pseudo-guessing parameter, and 1.7 is a scaling factor. In the absence of guessing, $c_i$will be set to 0 and Equation 2 will be reduced to the 2-parameter logistic (2PL) model. One example of an item without guessing is a short answer question scored dichotomously. The 1-parameter logistic (1PL) model further assumes the

discrimination parameter a_i to be constant across all items. Finally, when only the difficulty is modeled, the model reduces to the simplest form known as the Rasch (1960) model, written as

$$P(X_i = 1|\theta) = \frac{1}{1+\exp[-(\theta - b_i)]}. \quad (3)$$

Assumptions under all these IRT models include monotonicity, unidimensionality, and local independence. Monotonicity is met when $P(X_i = 1|\theta)$ is a non-decreasing function of θ. The unidimensionality assumptions require that all test items measure only one latent trait. The local independence assumption states that once θ is accounted for, responses to test items should be independent. More formally, let the vector of item response random variables be $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ and its realization be $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. Given the local independence assumption, the probability or likelihood of observing the response vector $\mathbf{x}$ can be expressed as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P\,(X_i = 1|\theta)^{x_i}[1 - P(X_i = 1|\theta)]^{1-x_i}. \quad (4)$$

Many parametric person-fit statistics make use of this likelihood function. For instance, the l_0 statistic introduced by Levine and Rubin (1979) is expressed as

$$l_0 = \sum_{i=1}^{k}[\,x_i \ln P_i + (1 - x_i)\ln(1 - P_i)], \quad (5)$$

where P_i is an abbreviation of $P(X_i = 1|\theta)$ in Equation 2. As shown, l_0 is the sum of the log-likelihood of the observed responses across the entire test for an examinee, which directly measures the fit of the data to the model. However, l_0 varies by θ and its sampling distribution is unknown, thus cannot be used for detecting the aberrant patterns conveniently. In order to overcome these limitations, Drasgow, Levine, and Williams (1985) standardized the l_0 and developed the l_z statistic, expressed as

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}}, \quad (6)$$

where $E(l_0)$ is the expected l_0, and the denominator is the standard error of l_0. The expectation can be computed as

$$E(l_0) = \sum_{i=1}^{k}[\, P_i \ln P_i + (1 - P_i)\ln(1 - P_i)], \tag{7}$$

and the variance can be derived by

$$Var(l_0) = \sum_{i=1}^{k} P_i\,(1 - P_i)\left[\ln \frac{P_i}{1-P_i}\right]^2. \tag{8}$$

The l_z was assumed to follow the standard normal distribution. Compared to eight other person-fit indices under the 2PL and 3PL models, Drasgow et al. (1987) found that l_z showed controlled Type I error and the highest overall detection rates of aberrant behavior across ability levels. However, others have argued that l_z is negatively skewed when the estimate of θ is used in place of the true θ (Molenaar & Hoijtink, 1990; Nering, 1995). In such a case, Type I error rates tend to be too conservative and power suffers accordingly. For these reasons, Snijders (2001) proposed l_z^*, which corrects the mean and variance of l_z when $\hat{\theta}$ is used in the calculation. According to Magis, Raiche, and Beland (2012), the l_z^* is calculated as

$$l_z^* = \frac{l_0(\hat{\theta}) - E[l_0(\hat{\theta})] + c_k(\hat{\theta}) + r_0(\hat{\theta})}{\sqrt{\widetilde{Var}[l_0(\hat{\theta})]}}, \tag{9}$$

where $\hat{\theta}$ is the estimate of θ, $l_0(\hat{\theta})$ is l_0 calculated using $\hat{\theta}$, and

$$\widetilde{Var}[l_0(\hat{\theta})] = \sum_{i=1}^{k} \widetilde{w_i}\,(\hat{\theta})^2 P_i(1 - P_i), \tag{10}$$

$$\widetilde{w_i}(\hat{\theta}) = w_i(\hat{\theta}) - c_k(\hat{\theta}) r_i(\hat{\theta}), \tag{11}$$

$$w_i(\hat{\theta}) = \ln \frac{P_i}{1-P_i}, \text{ and} \tag{12}$$

$$c_k(\hat{\theta}) = \frac{\sum_{i=1}^{k} P_i' w_i(\hat{\theta})}{\sum_{i=1}^{k} P_i' r_i(\hat{\theta})}, \tag{13}$$

where P_i' is the first derivative of P_i with respect to θ. The functions $r_0(\hat{\theta})$ and $r_i(\hat{\theta})$ depend on the estimation method of θ and the measurement model, where they must satisfy the equation

$$r_0(\hat{\theta}) + \sum_{i=1}^{k}[X_i - P_i]r_i(\hat{\theta}) = 0. \quad (14)$$

For example, if $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ, then

$$r_0(\hat{\theta}) = 0 \quad \text{and} \quad r_i(\hat{\theta}) = \frac{P_{i'}}{P_{i'}(1-P_{i'})}. \quad (15)$$

For the 2PL model using MLE of θ across test lengths, ability levels, and α conditions, empirical Type I error rates for l_z^* were closer to α than that for l_z (Snijders, 2001). The Type I error was recovered for all cases except for short tests (15 items or fewer) with the α levels lower than .05. de la Torre and Deng (2008) contended that Snijders (2001) corrected only the mean and variance of l_z, so the distribution of l_z^* was still negatively skewed, especially when the test was short. They proposed a method to define the sampling distribution through resampling and found improved Type I error rates that closely reflected the nominal level in all studied conditions. Other researchers have continued to improve and extend the l_z and the l_z^* statistic. For example, Sinharay (2016a) showed how to incorporate various estimation methods for θ in calculating l_z^*. Also, Sinharay (2016b) discussed resampling-based approaches to correcting the l_z^* that are generalizations of the method presented by de la Torre and Deng (2008). Meanwhile, l_z have been applied to less common IRT models (e.g., Lee, Stark, Chernyshenko, 2014) or subtests (Drasgow, Levine, & McLaughlin, 1991).

Overall, likelihood-based person-fit statistics have been well-studied and are often cited as the most popular person-fit statistics (e.g., de la Torre & Deng, 2008; Magis, Raiche, & Beland, 2012). These statistics are generally easy to compute. Their sampling distributions are well defined. They can handle missing data and perfectly correct and incorrect response patterns. One drawback, however, is that they are based on parameter estimates of the exact model whose validity is in question. Thus, researchers have proposed non-parametric person-fit statistics that do not rely on model estimates.

Non-Parametric Person-Fit Statistics

Non-parametric person-fit statistics are also called group-based person-fit statistics because they compare the examinee's responses to the other responses in the sample. Many of these statistics are based on the deterministic Guttman (1944, 1950) model. According to the Guttman model,

$$P(X_i = 1|\theta) = \begin{cases} 1 & \text{for } \theta \geq b_i \\ 0 & \text{for } \theta < b_i, \end{cases} \tag{16}$$

where b_i is the item difficulty parameter on the same scale as θ. Under this model, a Guttman error is committed when any items i and h that satisfies $b_i < b_h$ have the observed responses $X_i = 0$ and $X_h = 1$ because the response patterns do not conform to the model. Most non-parametric person-fit statistics based on the Guttman model use the equation below (Meijer, 2001). Let the items on a test of length k be arranged in descending ordered by the proportion of correct responses, $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_k$. Then, the standardized weighted proportion of Guttman errors is

$$B = \frac{\sum_{i=1}^{s} w_i - \sum_{i=1}^{k} X_i w_i}{\sum_{i=1}^{s} w_i - \sum_{i=k-s+1}^{k} w_i}, \tag{17}$$

where s is the sum of item scores and w_i is a weight that is defined by the particular person-fit statistic. Usually, $B = 0$ indicates perfect conformity to the Guttman model and $B = 1$ indicates complete model misfit.

For example, the Modified Caution Index (MCI; Harnisch & Linn, 1981) is obtained by setting $w_i = \pi_i$. The Caution Index (C; Sato, 1975) is obtained by setting $w_i = \pi_i$, then multiplying $\sum_{i=k-s+1}^{k} w_i$ by s and the other three terms by k. The U3 statistic (Flier, 1980, 1982) is obtained by setting $w_i = \ln[\pi_i/(1 - \pi_i)]$. Karabastos (2003) evaluated the performance of these statistics. The MCI, C, and U3 were among the best performing statistics in detecting

cheating, creative responding, lucky guessing, careless errors, and random responding. The best performing person-fit statistic in that study was also non-parametric, called the H^T statistic (Sijtsma, 1986). The H^T for examinee n is given as

$$H_n^T = \frac{\sum_{n \neq m} \beta_{nm} - \beta_n \beta_m}{\sum_{n \neq m} \max\{\beta_m(1-\beta_n), \beta_n(1-\beta_m)\}}, \quad (18)$$

where β_n and β_m are the proportion correct score for examinee n and m respectively, and β_{nm} is the proportion of items to which both examinees n and m answered correctly. H^T ranges from 1 to -1, where $H^T = 1$ indicates perfect fit to Guttman model. Similar to Karabastos (2003), Tendeiro and Meijer (2014) also found that the H^T was the best of seven person-fit indices in detecting aberrant behavior. The finding that the H^T was not correlated strongly with the sum score provides further evidence of its utility.

Based on these studies, non-parametric person-fit statistics frequently have higher power than their parametric peers. One possible reason is that parametric statistics often have deflated Type I error rates when estimated item parameters are used in the calculation (e.g., Tendeiro & Meijer; 2014), which inevitably reduces their power. On the other hand, non-parametric statistics have their own limitations. As they are generally based on cut-off values rather than sampling distributions, defining these cut values is not always easy and values used can vary by testing conditions, thus inconvenient for use in practice. Additionally, they often cannot handle perfect correct or incorrect response patterns. Presence of missing data can pose additional challenge.

Person Response Function

Parametric or non-parametric, the above person-fit statistics all suffer one major shortcoming: they are unable to reveal the possible cause of the aberrant behavior. Analyses using the person response function (PRF; Trabin & Weiss, 1983) may overcome this limitation. The PRF gives the probability of a correct answer for an individual with a fixed θ as a function

of item difficulty δ_i. The δ_i can be the b_i from the IRT models or other difficulty measures (Sijtsma & Meijer, 2001). The PRF is assumed to be a non-increasing function of δ_i. It can be further assumed that items have an invariant item ordering (Sijtsma & Junker, 1996), meaning the item response functions do not intersect. This assumption holds when all items fit the 1PL or the Rasch model, but may be violated under the 2PL and 3PL models. By comparing the observed and expected PRF, the analyst may be able to determine the pattern of person misfit in relation to the item difficulty.

To create the PRF, all k items are ordered from the easiest to the hardest, such that

$$\delta_1 \leq \delta_2 \leq \cdots \leq \delta_k, \tag{19}$$

where, under the invariant item ordering assumption, it holds that

$$P_1 \geq P_2 \geq \cdots \geq P_k, \tag{20}$$

for all θ, where P_i is an abbreviation of $P(X_i = 1|\theta)$. These items are divided into M non-overlapping subtests denoted as S_j $(S_1, S_2, \ldots, S_M)$, each containing t items. Then, $S_1 = \{1,2, \ldots, t\}, S_2 = \{t+1, \ldots, 2t\}, \ldots, S_M = \{k-t+1, \ldots, k\}$, and $M \times t = k$. The expected PRF is found by calculating the expected number of correct responses in subtest j, which is

$$t^{-1} \sum_{i \in S_j} P_i. \tag{21}$$

Also, the observed PRF is given by

$$t^{-1} \sum_{i \in S_j} X_i. \tag{22}$$

Therefore, the difference between the observed and expected PRF for subtest j is

$$D_j(\theta) = t^{-1} \sum_{i \in S_j} [X_i - P_i]. \tag{23}$$

Examples of the observed and expected PRF across six subtests for a hypothetical examinee with $\theta = 0$ are illustrated in Figure 1. If the examinee fits the model well, $D_j(\theta)$ will be near 0 for all subtests. However, in the presence of aberrant testing behaviors, say cheating

that allowed an examinee to obtain spuriously high scores, $D_j(\theta) \geq 0$ should be expected for most j. Also, careless mistakes can cause spuriously low scores, for which the PRF may show $D_j(\theta) \leq 0$ for most j. Another indication of person-misfit is a flat observed PRF relative to the expected PRF (Lumsden, 1977; Reise, 2000; Trabin & Weiss, 1983). A flat PRF can also be described as having $D_j(\theta) \leq D_u(\theta)$ where u is any subtest that contain items that are more difficult than those in j. Finally, the greatest absolute discrepancies between the observed and expected PRF for the spuriously high case usually occurs in the difficult subtests, while the opposite is true for the spuriously low situation.

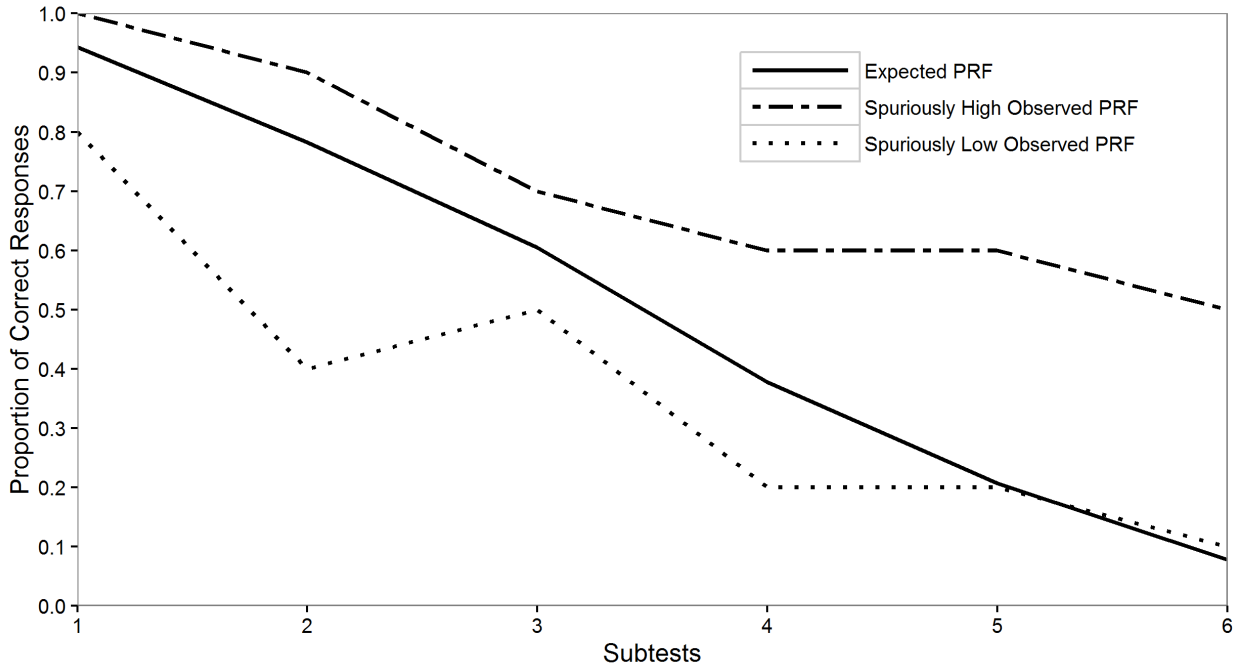

Figure 1. Example expected and observed person response functions (PRF) for an examinee with $\theta = 0$.

To investigate the usefulness of the PRF, Trabin and Weiss (1983) used data from a vocabulary test with 216 multiple-choice items among 151 examinees. The b_i from a 3PL model was used for δ_i. The test was divided into 9 subtests ($M = 9$), each containing 24 items ($t = 24$).

To statistically test the difference of the observed and expected PRF, a χ^2 goodness-of-fit test with $df = M - 2$ was conducted at $\alpha = .05$, calculated as

$$\chi^2 = \sum_{j=1}^{M} \frac{[\sum_{i \in S_j}(X_i - P_i)]^2}{\sum_{i \in S_j} P_i}. \quad (24)$$

Trabin and Weiss (1983) detected 15 misfitting examinees. Looking at the expected and observed PRFs for each individual allowed them to speculate the specific aberrant behavior that occurred. For example, "testwiseness" may describe a student skilled in guessing, which may manifest as high scores on very difficult items. Another PRF with low scores on easy items may indicate careless test-taking behavior..

Since then, researchers have attempted to estimate the slope of the PRF. Strandmark and Linn (1987) achieved this by adding to the typical 2PL model, a slope parameter that was allowed to vary by individual. On the other hand, Reise (2000) proposed that the multilevel logistic regression can be used to estimate the PRF slope by treating the items as nested within the individuals, and predicting the responses from δ_i and $\hat{\theta}$. Using the empirical Bayes estimation method, steeper PRF slopes (i.e., more negative) were associated with higher person-fit based on the l_z statistic (Pearson $r = -.65$). The procedure also allows inclusion of additional predictor variables that may aid the researcher in understanding the aberrant behavior.

More recently, Sijtsma and Meijer (2001) proposed a non-parametric statistic ρ based on the hypergeometric distribution to test the non-increasingness of the PRF slope. Although ρ was less powerful in detecting aberrance than the non-parametric person-fit statistic U3 (Van der Flier, 1982), ρ was still useful for understanding what sorts of aberrant behavior may have occurred. Emons, Sijtsma, and Meijer (2005) extended the research on PRFs by presenting a three-step non-parametric person-fit analysis. First, they used the U3 statistic to assess global person-fit of 1,641 children to four cognitive development tests, each containing 45 to 65 items.

They discussed the analysis of six particular cases who were determined deviant. The observed PRF for each individual was presented graphically using standard normal kernel smoothing. Areas of the PRF that showed visual evidence of increasingness were tested for non-increasingness using the G statistic (Emons, 2003). G is simply the total number of Guttman errors in the J items that fall in between the range of interest in the PRF, calculated as

$$G = \sum_{h=1}^{J} \sum_{i=1}^{h} (1 - X_i) X_h, \tag{25}$$

where $i \leq h$, and therefore $\delta_i \leq \delta_h$. Using the Wilcoxon's rank-sum distribution, $P(G \geq g | \sum_{i=1}^{J} x_i, J)$ was found and compared to $\alpha = .05$, where g is the realization of G and x_i is the realization of X_i. The authors identified that many of the observed increasing PRF slopes were in fact statistically significantly increasing. Based on the location of the increasingness in the PRF, the authors were able to speculate possible testing behaviors that caused the aberrance.

It is worth pointing out that testing for the non-increasingness of the PRF may fail to identify PRFs that are aberrant but not decreasing quickly enough. This limitation can be overcome by modeling the PRF as a non-increasing logistic curve and estimating the angle of the slope (e.g., Strandmark & Linn, 1987; Reise, 2000), but this approach may fail to identify important locations in the observed PRF that may actually be increasing. Lastly, a big challenge to most parametric PRF analyses is to accurately calculate the expected PRF because θ must be estimated in practice.

Robust Ability Estimation

A common ability estimation method under the IRT is the maximum likelihood estimation (MLE) method, which produces consistent, asymptotically efficient, and normally distributed estimates as long as there is a mix of both correct and incorrect responses (Birnbaum, 1968; Hambleton & Swaminathan, 1985). Let the vector of item response random variables be

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$ and its realization be $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, where $X_i = 1$ and $X_i = 0$ denotes a correct and incorrect response on item i, respectively. Given the local independence assumption, the log likelihood of a response vector $\mathbf{X}$ is

$$l = \sum_{i=1}^{k}[X_i \ln P_i + (1 - X_i)\ln(1 - P_i)] \quad (26)$$

where P_i is the probability of a correct response on item i based on a measurement model. The MLE of ability is simply the θ that maximize the likelihood or log-likelihood function. To find the maxima of equation 26, one can take the first derivative of the log-likelihood with respect to θ and setting it to zero, or

$$\sum_{i=1}^{k} \frac{dl_i}{d\theta} = \sum_{i=1}^{k} \left[\frac{X_i - P_i}{P_i(1-P_i)}\right] \frac{dP_i}{d\theta} = 0, \quad (27)$$

where l_i is the log-likelihood of item i. This equation can be solved by applying the Newton-Rapson algorithm. In case that $\mathbf{X}$ has been contaminated by an aberrant behavior, standard MLE $\hat{\theta}$ will not be able to account for its adverse effect.

When an examinee's response pattern is detected as aberrant, Smith (1985) listed the following possible actions:

1. Drop the score and retest the examinee.
2. Make a justification that the measurement error is small enough and report $\hat{\theta}$ as it is.
3. Report multiple $\hat{\theta}$s using model-fitting subtests.
4. Modify item responses and re-estimate θ.

Option 1 requires multiple sittings for the examinee; not practical in most testing conditions. Option 2 can be highly subjective. Moreover, measurement error is unlikely to be small for an aberrant response pattern. It is also hard to argue for using an estimate based on a wrong model. Option 3 not only increases measurement error due to the shortened test length but also makes scores incomparable. To sum up, options 1-3 are not ideal.

The fourth option looks most appealing. Given that most high stakes tests use a large number of items, it may be possible to modify or remove responses to some items so that the accuracy of $\hat{\theta}$ is satisfactory. One approach is to downweight the contributions of potentially aberrant responses to the likelihood function (Mislevy & Bock, 1982; Schuster & Yuan, 2011). The weighted maximum likelihood estimate is the value of θ that satisfies the equation

$$\sum_{i=1}^{k} g\,(r_i)\left(\frac{dl_i}{d\theta}\right) = 0, \tag{28}$$

where r_i is a residual value and $g(r_i)$ is the weight function resulting in values from 0 to 1. If $g(r_i) = 1$ for all i, then equations 27 and 28 will be equivalent. To extend the previous work by Waller (1974) and Wainer and Wright (1980), Mislevy and Bock (1982) introduced the biweight estimation method which modifies the likelihood function so that all items, regardless of the observed responses, are downweighted when the difficulty parameter of the item is far from the estimated ability. In a biweight estimate, $r_i = a_i(\theta - b_i)$ in a 2PL model and

$$g(r_i) = \begin{cases} \left[1 - \left(\frac{r_i}{B}\right)^2\right]^2 & \text{for } |r_i| \le B \\ 0 & \text{for } |r_i| > B, \end{cases} \tag{29}$$

where the tuning coefficient $B = 4$ was recommended by Mislevy and Bock. Decreasing B heightens the intensity of response downweighting. Their simulation study showed that the biweight estimate is less biased than the MLE for a variety of conditions when aberrant responses are present. However, Schuster and Yuan (2011) explained that the biweight estimation is not guaranteed to converge to an estimate when almost all responses are correct (or incorrect). This is because the biweight method allows zero weights, and sometimes the non-zero-weighted items become all incorrect or all correct. Therefore, they introduced the Huber weight estimate where r_i was kept the same but the weight function was modified to

$$g(r_i) = \begin{cases} 1 & \text{for } |r_i| \leq B \\ \frac{B}{|r_i|} & \text{for } |r_i| > B, \end{cases} \quad (30)$$

where the tuning coefficient $B = 1$ was recommended. Decreasing B heightens the intensity of response downweighting. Unlike the biweight approach, Huber weight estimation avoids placing a zero weight to any item. Schuster and Yuan (2011) showed in their simulation that both downweighting approaches limit the bias caused by aberrant responses compared to MLE, particularly for examinees with extreme $\theta's$. Huber weight and biweight methods are about equally effective.

These methods work because they limit the less informative item responses from affecting the estimate. For example, if the MLE $\hat{\theta}$ is 2, Huber weight and biweight methods heavily downweight the contribution of all easy items (e.g., $b_i < -1$) on $\hat{\theta}$. These items contain little information about this person's θ because correct responses are almost certain to happen; yet, incorrect responses due to aberrant behaviors can have a heavy influence on $\hat{\theta}$. For the non-aberrant items, downweighting will remove only a small amount of information, which may be compensated by the information gain from handling the aberrant responses. Therefore, these downweighting estimation methods attempt to limit the aberrant responses from influencing $\hat{\theta}$, while minimizing the detrimental effects of downweighting the valid responses.

Both Reise (1995) and Meijer and Nering (1997) attempted to use the biweight estimation method to improve person-misfit detection using the l_z statistic (Drasgow et al., 1985). Both studies compared the biweight method with MLE and expected a-posteriori (EAP) methods. EAP is a Bayesian estimation method that combines prior information with the likelihood to form a posterior distribution of θ (Hambleton & Swaminathan, 1985). The posterior density $f(\theta|\mathbf{x})$ can be expressed as

$$f(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)f(\theta), \quad (31)$$

where $L(\mathbf{x}|\theta)$ is the likelihood function and $f(\theta)$ is the prior of θ. The mean of the posterior distribution is the EAP estimate, while the mode is the maximum a-posteriori (MAP) estimate. If $f(\theta)$ is a constant, the posterior is proportional to the likelihood and the MLE and MAP estimates will be equivalent. While the prior distribution depends on the prior knowledge one has on the target parameter, $N(0,1)$ is often used due to its proximity to the default scale of the trait implemented in many computer programs. When an appropriate informative prior is used, EAP and MAP should be more accurate than MLE.

Reise (1995) found that when the biweight estimate was used instead of MLE or EAP, the change in the power of the l_z statistic was minimal. Meijer and Nering (1997) argued that Reise's disappointing results may be due to the fact that aberrant responses were simulated by setting the probability of a correct response for aberrant items as $P^* = .5$. Such a manipulation has minimal effects on examinees with θ near 0 (i.e., P is close to .5 anyway) and only moderate effects on those with extreme θ's. Meijer and Nering hence simulated aberrant responses under three conditions: $P^* = .5$ for all θ, $P^* = 1$ among $\theta \leq 0$, and $P^* = .2$ among $\theta \geq 0$. They found evidence supporting their hypotheses and the detection rates improved particularly for individuals with an extreme θ (e.g., $\theta = 2$ and $\theta = -2$). However, if a more accurate $\hat{\theta}$ can be found, the power of person-fit statistics could improve further (Reise, 2000).

To further the area, Maeda and Zhang (2017b) recently developed the biweight-MAP (BMAP) and Huber weight-MAP (HMAP), which combines the elements of robust estimation with Bayesian methodology. BMAP and HMAP are the modes of the posterior distribution with a downweighted likelihood using the biweight and Huber weight, respectively. The procedure aims to use the information from the prior distribution to compensate for the information loss

due to aberrant responses, and resist the detrimental effects of downweighting the fitting items by either weighting method.

BMAP and HMAP can be calculated using the Newton-Rapson algorithm. In essence, for the 2PL model with a $N(\mu, \sigma^2)$ prior, the improved estimate $\hat{\theta}^v$ for the v^{th} iteration is

$$\hat{\theta}^v = \hat{\theta}^{v-1} - \frac{\sum[g(r_i)^v a_i(X_i - \hat{P}_i)] - \frac{\theta-\mu}{\sigma^2}}{-\sum[g(r_i)^v a_i^2 \hat{P}_i(1-\hat{P}_i)] - \frac{1}{\sigma^2}}. \quad (32)$$

where the numerators and the denominators are the first and second derivatives of the posterior probability with respect to $\hat{\theta}$, respectively. When calculating the BMAP, the weight $g(r_i)$ is defined as the biweight (Equation 29), while Huber weights (Equation 30) are used for HMAP. The algorithm usually converges in a few iterations. For more details of these two new estimates, refer to Maeda and Zhang (2017b). A Monte Carlo simulation showed that, out of all the studied robust and non-robust ability estimates (i.e., MLE, MAP, biweight, BMAP, and HMAP), BMAP had the smallest root-mean-squared-error under the 10% and 20% aberrant items conditions. Despite these promising results, BMAP may still be improved by taking into account the model-fit of the observed responses. By downweighting responses that are particularly misfitting, it may be possible to further correct the bias due to the aberrant responses.

CHAPTER 3: PROPOSED ROBUST ABILITY ESTIMATION METHOD

In this chapter, a new robust ability estimation method for aberrant examinees is presented. The procedure iteratively identifies and downweights potentially aberrant responses based on low item information and poor model fit until the response vector fits the model sufficiently well. Identification of aberrant responses is based on the person response function (Trabin & Weiss, 1983), while uninformative items receive downweighting based on the BMAP (Maeda & Zhang, 2017b). This method is named the downweighting of aberrant responses estimation (DARE) method. DARE consists of the following six steps.

Step 1: Person-Fit Testing and Initial Ability Estimation

First, global person-fit is assessed to identify examinees that require robust ability estimation. A good person-fit statistic will detect as many aberrant but as few non-aberrant examinees as possible. One example is the l_z^* based on the MAP estimate using the .01 significance level (Maeda & Zhang, 2017b). Once potentially aberrant examinees are detected, the initial ability estimates can be obtained by the BMAP method using the tuning coefficient $B = 4$ as recommended by previous studies (Maeda & Zhang, 2017b; Mislevy & Bock, 1982).

Step 2: Assembling Subtests

Similar to constructing the observed PRF (Trabin & Weiss, 1983), the test is divided into subtests. Let the vector of a response pattern be $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ and its realization be $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, where $X_i = 1$ and $X_i = 0$ denotes a correct and incorrect response on item i, respectively. Order all k items from the easiest to hardest, using the initial BMAP estimate $\hat{\theta}$, shown as

$$\delta_1 \leq \delta_2 \leq \cdots \leq \delta_k, \tag{33}$$

where δ_i is the conditional probability $P(X_i = 0|\hat{\theta})$. Alternatively, δ_i can also be the proportion of incorrect responses in the sample or the estimated item difficulty from the IRT model. However, such choices would require the assumption of invariant item ordering (Sijtsma & Junker, 1996). By conditioning δ_i on $\hat{\theta}$, the need for assuming invariant item ordering is partly avoided, but not completely due to the error in θ estimation.

Test items are assigned to M non-overlapping subtests in order denoted as S_j $(S_1, S_2, \dots, S_M)$, each containing t_j items, thus $S_1 = \{1,2, \dots, t_1\}, S_2 = \{t_1 + 1, \dots, t_1 + t_2\}, \dots, S_M = \{k - t_M + 1, \dots, k\}$. The average proportion of correct responses for every S_j is expected to be non-increasing, shown as

$$t_1^{-1} \sum_{i \in S_1} P_{i \in S_1} \geq t_2^{-1} \sum_{i \in S_2} P_{i \in S_2} \geq \cdots \geq t_M^{-1} \sum_{i \in S_M} P_{i \in S_M}, \quad (34)$$

where $P_{i \in S_j}$ is the abbreviation of $P\left(X_{i \in S_j} = 1 \middle| \hat{\theta}\right)$. The main reason to assemble subtests by P_i is that aberrant behavior often affects items based on the item difficulty. As not all items are affected, grouping items by difficulty will increase the power to detect the aberrant responses.

The following points should be considered when forming subtests. First, the difficulty level of items in a subtest should be nearly homogeneous. Second, subtests should be large enough so that there is adequate power to assess person-fit. Finally, while it is convenient to have subtests of equal size, this is not required. Given these considerations, pilot results show that $M = 3$ subtests with equal numbers of items is a good balance between subtest length and item difficulty homogeneity in most tests with 20 to 60 items.

Step 3: Identifying the Least Fitting Subtest

This step identifies the least fitting subtest using local person-fit statistics. To quantify the local fit of the responses, the l_{zw} is calculated for each subtest independently using the BMAP estimate $\hat{\theta}$. In this step, $\hat{\theta}$ is based on all the responses in the test rather than the responses in

each subtest because such estimates can be highly unreliable due to the short test length. The l_{zw} is a weighted version of the l_z, where the weight w_i for item i is inserted into Equations 5 to 8 in the calculation:

$$l_{zw} = \frac{l_{0w} - E(l_{0w})}{\sqrt{Var(l_{0w})}}, \tag{35}$$

where the l_{0w} is the weighted log-likelihood, $E(l_{0w})$ is its expected value, and $Var(l_{0w})$ is its variance. Initially, $w_i = 1$ for all items, but may decrease to as low as 0 in later iterations within DARE. The weighted log-likelihood is found simply by multiplying the log-likelihood for each item by w_i:

$$l_{0w} = \sum_{i \in S_j} w_i [\, x_i \ln P_i + (1 - x_i) \ln(1 - P_i)]. \tag{36}$$

Therefore, the modification to the expected value equation (Equation 7) is straight forward:

$$E(l_{0w}) = \sum_{i \in S_j} w_i [\, P_i \ln P_i + (1 - P_i) \ln(1 - P_i)]. \tag{37}$$

Finally, $Var(l_{0w})$ is found from the fact that it is equivalent to $E[(l_{0w} - E(l_{0w}))^2]$, which can be reduced to

$$Var(l_{0w}) = \sum_{i \in S_j} w_i^2 P_i \,(1 - P_i) \left[\ln \frac{P_i}{1 - P_i}\right]^2. \tag{38}$$

Like the l_z, the l_{zw} is assumed to follow the standard normal distribution when the parameters are known and there are many items. However, unlike the l_z, the l_{zw} is able to take into account the weights calculated in DARE. Also, as perfectly correct and incorrect response patterns are likely to appear in subtests, parametric person-fit statistics such as the l_{zw} may work better than some non-parametric methods.

A highly negative l_{zw} indicates that the subtest has a flat observed PRF relative to the expected PRF, representing person misfit (Lumsden, 1977; Reise, 2000; Trabin & Weiss, 1983). The l_{zw} for a subtest will be highly negative if the observed PRF is higher than expected when

the expected PRF is less than .5, or when the observed PRF is lower than expected when the expected PRF is more than .5. The subtest with the lowest l_{zw} is the least fitting subtest, labeled as S_A.

Step 4: Downweight One Misfitting Response

Theoretically, all responses in S_A that meet the condition $P\left(X_i = x_i \middle| \hat{\theta}\right) < .5$ and $w_i > 0$ can be considered as potentially aberrant and may have contributed to person-misfit. The next step is to downweight one of them. The item with the lowest a parameter estimate is selected and its w_i is simply subtracted by 0.1. This item is a reasonable choice because it has the most stable item response function across θ (i.e., flat), which is important given that $P\left(X_i = x_i \middle| \hat{\theta}\right)$ may be an inaccurate estimate of $P(X_i = x_i|\theta)$. In addition, this item also contains low discriminating power; downweighting it has a small drawback in case that it is actually a non-aberrant response. The item with the lowest $P\left(X_i = x_i \middle| \hat{\theta}\right)$ is potentially a good candidate for downweighting as well, but this criteria is not used because the biweights in the BMAP procedure already places the lowest weights on these items.

Step 5: Evaluate Ability Estimate and Person-Fit

The next step is to re-evaluate the ability estimate and global person-fit using the new weights. First, the BMAP is re-calculated with modifications to the biweights. The tuning coefficient is set at $B = 5$, which is larger than the recommended $B = 4$ (Mislevy & Bock, 1982). This downweights slightly less intensely than the typical biweight method, leaving room for further downweighting from w_i. Further, the biweights are multiplied by w_i. Therefore, the weight function $g(r_i)$ in Equation 29 is modified to

$$g(r_i, w_i) = \begin{cases} w_i \left[1 - \left(\frac{r_i}{5}\right)^2\right]^2 & \text{for } |r_i| \leq 5 \\ 0 & \text{for } |r_i| > 5. \end{cases} \quad (39)$$

The BMAP is re-calculated using the modified biweights $g(r_i, w_i)$, and the new estimate is labeled as $\hat{\theta}_v$. Using the new ability estimate $\hat{\theta}_v$ and the updated weights w_i, global person-fit is evaluated by calculating l_{zw} using all responses (i.e., with all subtests combined). To be clear, the modified biweights $g(r_i, w_i)$ are never used in the person-fit analysis.

Step 6: Deciding the Convergence

Steps 2 to 5 are iterated until global person-fit is satisfactory. Every iteration, one response is downweighted by 0.1 and $\hat{\theta}$ is updated and used to assess the person-fit. Once the global person-fit statistic satisfies the pre-determined cutoff value, such as $l_{zw} > -1.645$, the algorithm stops. Basing the convergence on a global person-fit statistic is advantageous because the downweighting should intensify as the severity of the aberrant behavior increases. In contrast, this severity is independent of the weights used in Huber weight (Schuster & Yuan, 2011), biweight (Mislevy & Bock, 1982), HMAP, and BMAP methods (Maeda & Zhang, 2017b). Once the response pattern is sufficiently fitting, $\hat{\theta}_v$ from the final iteration is the final ability estimate of that person.

CHAPTER 4: METHODS

A Monte Carlo simulation was conducted to examine the effectiveness of DARE in improving ability estimation for misfitting persons. The study design included conditions commonly examined in the person-fit literature (Rupp, 2013). Design variables were test length, percentage of aberrant examinees, percentage of aberrant items, and type of aberrant behaviors. These design variables were not fully crossed. Instead, the study conditions were made up of four components (see Table 1).

Table 1.
Study conditions

Design Variables	Non-Aberrant Conditions	Aberrant Conditions	Supplemental Conditions 1	Supplemental Conditions 2
Test Length	20, 40, 60	20, 40, 60	60	60
Aberrant Behavior	None	SH, SL, Mixed	Mixed	Mixed
Aberrant Examinees	0%	10%, 30%	10%, 30%	30%
Aberrant Items	0%	10%, 20%, 30%	50%	10%, 20%, 30%
Variance of Item Difficulty	1	1	1	4
Number of Conditions	3	54	2	3

Test Characteristics

Currently, the BMAP method has been studied only with 60-item tests (Maeda & Zhang, 2017b). This was extended to 20, 40, and 60 item tests in the current study. Item discrimination (a_i) parameters were randomly sampled from a $lnN(0.4, 0.5)$ distribution truncated within the interval [0.6, 3], while the difficulty (b_i) parameters were randomly sampled from a $N(0, 1)$ distribution truncated within the interval [-3, 3]. These specifications ensured all test items had adequate discrimination and covered a well-spread difficulty continuum. In theory, since very difficult and very easy items are highly informative about the examinee's aberrant behavior

(Reise & Due, 1991), the robust methods should perform better when the item difficulty distribution is flatter than centrally focused. To explore this matter, supplemental conditions were added where the difficulty parameters were sampled from a $N(0, 4)$ distribution truncated within the interval [-3, 3]. The a_i and b_i were generated independent of each other, and sampled independently for each item in each replication. Therefore, the general test structure was fixed but different items were used in each replication in order to increase the generalizability of the findings.

Aberrant Response Behaviors

The percentage of examinees with aberrant behaviors in each data set were simulated at three levels: 0%, 10%, or 30%. This percentage is a critical factor in influencing the accuracy of item parameter estimates and the aberrance detection rate (Rupp, 2013). For each aberrant examinee, the percentage of items suffering from aberrance was also studied at three levels: 10%, 20%, and 30%. This percentage signifies the severity of aberrant behavior. Additionally, to examine the extremely severe conditions, two 50% aberrant items conditions were added. Under these extreme conditions, as there were equal numbers of aberrant and non-aberrant responses, all robust estimation methods may fail and accurate ability estimation may be impossible.

Different from previous studies that targeted specific behaviors such as cheating or creative responding, this study investigated three types of general aberrant response patterns: spuriously low (SL), spuriously high (SH), and mixed (both SL and SH; Rupp, 2013). SL results from behaviors that lead to spuriously low test scores, such as misreading instructions, lack of motivation, and fatigue. To simulate SL, correct answers were changed to incorrect with a .8 probability, synonymous to random guessing on a 5-option multiple-choice item (.2 probability of a correct response), regardless of the ability level. SH represents aberrant behaviors that result

in increased test scores, such as cheating, possessing pre-knowledge of the answers, and excessive lucky guessing. To simulate SH, incorrect responses were randomly changed to correct. Finally, data sets in the mixed condition contained equal numbers of SL and SH examinees. Examinees that received aberrant responses were selected independently of examinee ability. As an exception, examinees were not simulated as aberrant if they started with fewer correct or incorrect model-fitting responses than the targeted number of aberrant responses. For instance, in a 20-item test, examinees with a total score lower than six were not selected in simulating 30% SL aberrant behavior.

Data Generation and Model Estimation

The sample size was fixed at 1,000 for all conditions. This is the most commonly studied sample size in the person-fit literature (Rupp, 2013). Additionally, it should be large enough to provide accurate parameter estimates for the 2PL model (Morizot, Ainsworth, & Reise, 2007). The person ability parameter θ was randomly sampled from a $N(0,1)$ distribution. Initial item responses were generated based on the 2PL model. These responses were modified for aberrant examinees according to their assigned conditions. MULTILOG 7.0 (Thissen, 1991) was used for item parameter estimation. To achieve stable results, 1,000 replications were conducted for every condition.

Person-Fit Testing

Two person-fit statistics were used to assess global person-fit using the PerFit package (Tendeiro, 2015) in R (R Core Team, 2015). The l_z^* statistic (Snijders, 2001) was included in the study because of its high performance and popularity (e.g., Magis et al., 2012). The l_z^* was calculated based on the MAP estimate of θ. The H^T statistic (Sijtsma, 1986) was also included for its high power (Karabastos, 2003; Tendeiro & Meijer, 2014). The cutoff value for H^T was

calculated in every simulation replication based on 10,000 responses generated using the estimated item parameters. Response vectors with all incorrect and all correct scores were automatically labeled as fitting because the H^T cannot be calculated for these examinees. The Type I error rates were calculated as the rejection rate for the examinees simulated without aberrance, while the power rate was from those with aberrance.

It is not always easy to determine the alpha level for the person fit statistics in robust estimation. Robust ability estimation aims to improve the ability estimates for examinees with aberrant testing behaviors. Meanwhile, it can hurt estimation for examinees without aberrant behaviors due to the information loss resulting from downweighting their responses. One way to achieve equilibrium in this situation is by adjusting the α level. Maeda and Zhang (2017b) proposed the .01 α level as the most appropriate threshold for the BMAP method. At this level, enough true positives are detected while false positives are limited. They also found that examinees detected at this threshold benefit from using the BMAP over the standard MAP while those above the cutoff do not. As Maeda and Zhang's (2017b) study was limited to tests with 60 items, this study revisited this issue by examining cutoff values at α of .05, .025, and .01 for all three test lengths.

It is important to point out that in the context of robust estimation, the controlled Type I error and high power, while desirable, are not necessarily indications of ideal performance for the person-fit statistics. More important was whether the detected individuals will benefit from applying DARE. To answer that question, the estimation accuracy of MAP, BMAP, and DARE was compared graphically across the percentiles of H^T and l_z^* null distributions.

Ability Estimation for Misfitting Examinees

Examinees detected as misfitting had their θ re-estimated using five methods: MLE, MAP, biweight, BMAP, and DARE. Note that Huber weight (Schuster & Yuan, 2011) and HMAP (Maeda & Zhang, 2017b) were not used due to their similarity to biweight and BMAP. All Bayesian methods were conducted using the informative $N(0,1)$ prior, as typically done in practice (Hambleton & Swaminathan, 1985). As recommended, tuning coefficients for biweight and BMAP were set at $B = 4$ (Mislevy & Bock, 1982). For DARE, the number of the subtests was fixed at three. The number of items in subtests 1 and 3 were a third of the test length rounded to the nearest whole number, and the rest of the items were placed in subtest 2. The global person-fit convergence criterion was fixed at $l_{zw} > -1.645$, which is based on a one-tailed .05 α level, commonly used for evaluating person-fit. Note that this criterion does not need to coincide with the one used in the initial person-fit analysis. From one theoretical perspective, a criterion as high as $l_{zw} \geq 0$ could be reasonable because this indicates average fit. However, such a liberal criterion was not used due to its tendency to downweight more non-aberrant responses. All estimates were bounded within the interval [-4, 4] to prevent extreme values. Ability estimation was programmed in R, using a portion of the code written by Maeda and Zhang (2017b), Schuster and Yuan (2011), and Rose (2010).

Dependent Variables

To evaluate the ability estimation accuracy, two statistics were used: root-mean-squared-error (RMSE) and bias. RMSE was calculated as

$$\sqrt{E[(\hat{\theta} - \theta)^2]}, \tag{40}$$

while bias was calculated as

$$E\left(\hat{\theta} - \theta\right). \tag{41}$$

For perspective, when estimation is perfect (i.e., $\theta = \hat{\theta}$ for every person), RMSE is 0. On the other hand, if $\hat{\theta}$ is the average of θ for every examinee (i.e., an extremely poor estimate), RMSE is simply the standard deviation of the true θ distribution, or 1 in the current study. Therefore, estimates are not at all useful when RMSE exceeds 1 in this study. In contrast, as bias is directional, a bias of 0 does not mean $\theta = \hat{\theta}$ for every person. Instead, it shows that the average of $\hat{\theta}$ equals the average of θ. Based on these considerations, low RMSE is important for interpreting individual estimates while low absolute bias is more useful for group-level estimation.

RMSE and bias were computed for the whole sample for all conditions as well as for nine equally spaced groups of θ between -2.0 and 2.0 for some conditions. Ability group assignment proceeded by rounding examinee θ to the nearest .5. The exceptions were that the first group included all examinees with $\theta < -1.75$ while the last group included all examinees with $\theta > 1.75$. As the detection rates of l_z^* are extremely dependent on θ (de la Torre & Deng, 2008), examining the results both marginal to and conditioned on θ can be highly valuable.

In order to further understand the weights w_i applied in DARE, two additional statistics were calculated for all conditions. Average weights on aberrant items for a single examinee was calculated as

$$\sum_{i \in A} \frac{w_i}{a}, \tag{42}$$

where w_i is the DARE weights for item i, A is the set of aberrant items, and a is the number of aberrant items. Also, average weights on non-aberrant items was defined as

$$\sum_{i \notin A} \frac{w_i}{k-a}, \tag{43}$$

where k is the total number of items. Average weights on aberrant items were calculated only for the aberrant examinees, while the average weights on non-aberrant items were calculated for all

examinees. These statistics were averaged across all examinees in each condition, separated by whether the examinee was aberrant. The values ranged from 1 to 0, where 1 indicated that no downweighting happened to the item while 0 indicated that the item was removed from the likelihood function. If the observed effectiveness of DARE coincides with its theoretical justification, the above statistics should show that the aberrant items have received more downweighting (lower weights) than the non-aberrant items among the aberrant examinees.

CHAPTER 5: RESULTS

Detection Rates of Aberrant Response Patterns

The H^T person-fit statistic for the 20-item tests controlled Type I error rates when aberrant behaviors were not present in the data (see Table 2). Type I error gradually deflated as the percentage of aberrant examinees and items increased. Overall, power was low, ranging from .021 to .275. Power increased with the increase of aberrant items but decreased with the increase of aberrant examinees. As expected, power was lower at the lower α levels and the drop was considerable. Finally, the detection rates were slightly lower for the SL aberrant behavior type compared to the SH and mixed conditions.

Table 2.
H^T Type I Error and Power (20 Items)

AB	AE	AI	α=.01		α=.025		α=.05	
			Type I	Power	Type I	Power	Type I	Power
None	0%	0%	.010		.025		.051	
SH	10%	10%	.009	.029	.022	.071	.044	.131
		20%	.008	.047	.020	.116	.040	.203
		30%	.007	.068	.019	.166	.038	.275
	30%	10%	.008	.022	.018	.053	.035	.102
		20%	.006	.027	.015	.073	.028	.140
		30%	.005	.032	.012	.090	.024	.173
SL	10%	10%	.009	.027	.023	.063	.046	.115
		20%	.009	.040	.021	.100	.043	.176
		30%	.008	.059	.020	.142	.040	.242
	30%	10%	.008	.021	.019	.051	.037	.095
		20%	.007	.026	.016	.069	.032	.130
		30%	.006	.032	.014	.087	.027	.163
Mix	10%	10%	.009	.029	.023	.068	.045	.125
		20%	.009	.044	.021	.108	.042	.190
		30%	.008	.064	.020	.152	.039	.253
	30%	10%	.008	.021	.019	.052	.036	.099
		20%	.007	.026	.015	.069	.030	.133
		30%	.005	.032	.013	.091	.025	.170

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

As shown in Table 3, the l_z^* person-fit statistic showed similar patterns. Type I error rates were better controlled when aberrant behaviors were not present, but deflated more dramatically than H^T as the percentage of aberrant examinees and items increased. The power patterns were also similar. The values ranged from .041 to .302, slightly but consistently higher than that of H^T for the corresponding conditions.

Table 3.
l_z^* Type I Error and Power (20 Items)

AB	AE	AI	α=.01		α=.025		α=.05	
			Type I	Power	Type I	Power	Type I	Power
None	0%	0%	.017		.031		.051	
SH	10%	10%	.012	.075	.024	.115	.041	.164
		20%	.010	.133	.021	.188	.037	.246
		30%	.010	.179	.020	.240	.035	.302
	30%	10%	.008	.045	.016	.078	.029	.118
		20%	.005	.077	.011	.120	.021	.170
		30%	.004	.098	.010	.145	.019	.198
SL	10%	10%	.013	.064	.025	.099	.043	.141
		20%	.011	.113	.022	.164	.038	.222
		30%	.010	.160	.020	.221	.035	.286
	30%	10%	.009	.041	.018	.071	.032	.107
		20%	.006	.065	.013	.105	.024	.151
		30%	.004	.086	.010	.133	.019	.188
Mix	10%	10%	.012	.071	.024	.111	.042	.156
		20%	.010	.129	.021	.182	.037	.240
		30%	.009	.178	.019	.241	.034	.306
	30%	10%	.007	.047	.016	.078	.029	.118
		20%	.005	.075	.010	.117	.020	.168
		30%	.003	.100	.008	.150	.016	.206

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

Results for the 40-item and 60-item conditions are presented in Tables 4 to 7. While the patterns observed for the 20-item tests persisted, increased test length decreased Type I error

rates and increased power. The l_z^* continued to show higher power than H^T. The pattern of lower power for SL persisted for these longer tests.

Table 4.
H^T Type I Error and Power (40 Items)

AB	AE	AI	α=.01		α=.025		α=.05	
			Type I	Power	Type I	Power	Type I	Power
None	0%	0%	.010		.025		.051	
SH	10%	10%	.009	.045	.022	.104	.043	.181
		20%	.008	.095	.019	.205	.038	.319
		30%	.007	.143	.017	.293	.034	.427
	30%	10%	.007	.032	.017	.076	.032	.136
		20%	.005	.046	.012	.118	.023	.209
		30%	.004	.057	.009	.158	.017	.280
SL	10%	10%	.010	.038	.023	.089	.045	.156
		20%	.009	.074	.020	.164	.040	.266
		30%	.008	.120	.019	.251	.037	.378
	30%	10%	.008	.028	.018	.067	.035	.121
		20%	.007	.045	.014	.108	.027	.189
		30%	.005	.064	.012	.154	.022	.259
Mix	10%	10%	.009	.042	.022	.096	.044	.168
		20%	.008	.085	.020	.185	.039	.293
		30%	.007	.131	.018	.270	.036	.396
	30%	10%	.008	.031	.017	.073	.033	.131
		20%	.006	.045	.013	.112	.025	.198
		30%	.004	.057	.010	.151	.019	.259

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

Table 5.
l_z^* Type I Error and Power (40 Items)

AB	AE	AI	α=.01		α=.025		α=.05	
			Type I	Power	Type I	Power	Type I	Power
None	0%	0%	.015		.030		.051	
SH	10%	10%	.010	.112	.022	.167	.039	.227
		20%	.009	.225	.018	.296	.034	.368
		30%	.008	.294	.017	.368	.031	.438
	30%	10%	.006	.066	.013	.109	.025	.159
		20%	.003	.124	.008	.181	.017	.243
		30%	.003	.168	.007	.229	.014	.293
SL	10%	10%	.011	.091	.023	.140	.041	.196
		20%	.009	.186	.019	.255	.035	.326
		30%	.008	.269	.017	.347	.031	.424
	30%	10%	.007	.056	.015	.094	.028	.140
		20%	.004	.104	.009	.159	.019	.220
		30%	.003	.146	.007	.210	.014	.279
Mix	10%	10%	.011	.105	.022	.157	.039	.215
		20%	.008	.216	.018	.288	.033	.361
		30%	.007	.302	.016	.382	.030	.457
	30%	10%	.005	.069	.012	.109	.024	.159
		20%	.003	.123	.007	.181	.015	.244
		30%	.002	.170	.005	.238	.011	.309

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

Table 6.
H^T Type I Error and Power (60 Items)

σ^2	AB	AE	AI	α=.01		α=.025		α=.05	
				Type I	Power	Type I	Power	Type I	Power
1	None	0%	0%	.010		.026		.052	
	SH	10%	10%	.009	.065	.022	.139	.043	.228
			20%	.008	.145	.018	.281	.037	.408
			30%	.006	.231	.016	.407	.032	.540
		30%	10%	.007	.043	.015	.096	.029	.167
			20%	.005	.075	.011	.171	.020	.282
			30%	.003	.088	.007	.228	.015	.376
	SL	10%	10%	.009	.053	.022	.115	.044	.195
			20%	.008	.115	.020	.229	.039	.345
			30%	.008	.195	.018	.352	.035	.488
		30%	10%	.008	.038	.017	.084	.032	.147
			20%	.006	.068	.013	.148	.024	.243
			30%	.005	.107	.011	.222	.019	.345
	Mix	10%	10%	.009	.058	.022	.127	.043	.209
			20%	.008	.135	.019	.258	.038	.376
			30%	.007	.212	.017	.373	.034	.501
			50%	.006	.384	.014	.570	.029	.680
		30%	10%	.007	.042	.016	.093	.031	.159
			20%	.005	.071	.011	.159	.021	.259
			30%	.004	.093	.009	.216	.016	.340
			50%	.002	.155	.005	.346	.010	.499
4	Mix	30%	10%	.004	.091	.009	.163	.019	.245
			20%	.002	.169	.004	.283	.009	.389
			30%	.001	.191	.003	.356	.006	.481

Note. σ^2=variance of the item difficulty parameter; AE=aberrant examinees; AI=aberrant items; AB=aberrant behaviors; SL=spuriously low; SH=spuriously high

Table 7.
l_z^* Type I Error and Power (60 Items)

σ^2	AB	AE	AI	α=.01		α=.025		α=.05	
				Type I	Power	Type I	Power	Type I	Power
1	None	0%	0%	.015		.029		.051	
	SH	10%	10%	.010	.149	.021	.214	.038	.285
			20%	.007	.300	.017	.378	.032	.453
			30%	.007	.382	.015	.457	.029	.525
		30%	10%	.005	.086	.011	.136	.022	.195
			20%	.002	.171	.006	.237	.013	.307
			30%	.002	.225	.005	.293	.011	.360
	SL	10%	10%	.010	.119	.022	.176	.039	.243
			20%	.008	.250	.017	.330	.033	.407
			30%	.007	.359	.015	.445	.029	.525
		30%	10%	.006	.071	.013	.116	.025	.170
			20%	.003	.141	.007	.204	.015	.273
			30%	.002	.201	.005	.274	.011	.351
	Mix	10%	10%	.010	.138	.021	.201	.038	.267
			20%	.007	.290	.016	.370	.031	.448
			30%	.006	.402	.014	.484	.027	.556
			50%	.005	.529	.012	.605	.024	.668
		30%	10%	.004	.087	.011	.136	.022	.193
			20%	.002	.169	.006	.237	.012	.309
			30%	.001	.237	.004	.315	.008	.393
			50%	.001	.346	.002	.430	.006	.507
4	Mix	30%	10%	.003	.139	.006	.202	.013	.270
			20%	.001	.269	.003	.348	.006	.425
			30%	.001	.355	.002	.431	.004	.501

Note. AB=aberrant behaviors; σ^2=variance of the item difficulty parameter; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

In Tables 6 and 7, the exploratory 50% aberrant item conditions on the 60-item tests are presented. Results show that these conditions had the highest power rates and the lowest Type I error rates for both H^T and l_z^*. Additionally, tests with more extreme item difficulties where the difficulty parameters were sampled from $N(0,4)$ instead of $N(0,1)$ showed considerable power increases while Type I error was deflated further. Overall, the person-fit analysis show that Type I error rates tend to be deflated for aberrant response patterns and l_z^* has consistently higher

power than H^T, although power was rarely over 0.5 for the majority of the conditions. Given the deflated Type I error, observed power may be underestimated. How the detected individuals benefited from applying DARE is explored in the next section.

Appropriate α Level for Robust Ability Estimation

To determine the optimal α level for person-fit detection, the estimation bias and RMSE of DARE, BMAP, and MAP were examined across the tails of the H^T and l_z^* null distributions. Bias and RMSE were calculated independently for each study condition, which were aggregated using boxplots and means. In Figure 2, bias is shown by the type of aberrant behavior and the null distribution percentile. As expected, the SH condition always showed positive bias, SL condition showed negative bias, and the mixed condition showed almost no bias. The quartiles of bias were more extreme for worse-fitting examinees. Most notably, on average, the DARE and BMAP methods showed less bias compared to MAP only when examinees met the < 1 person-fit percentile criterion.

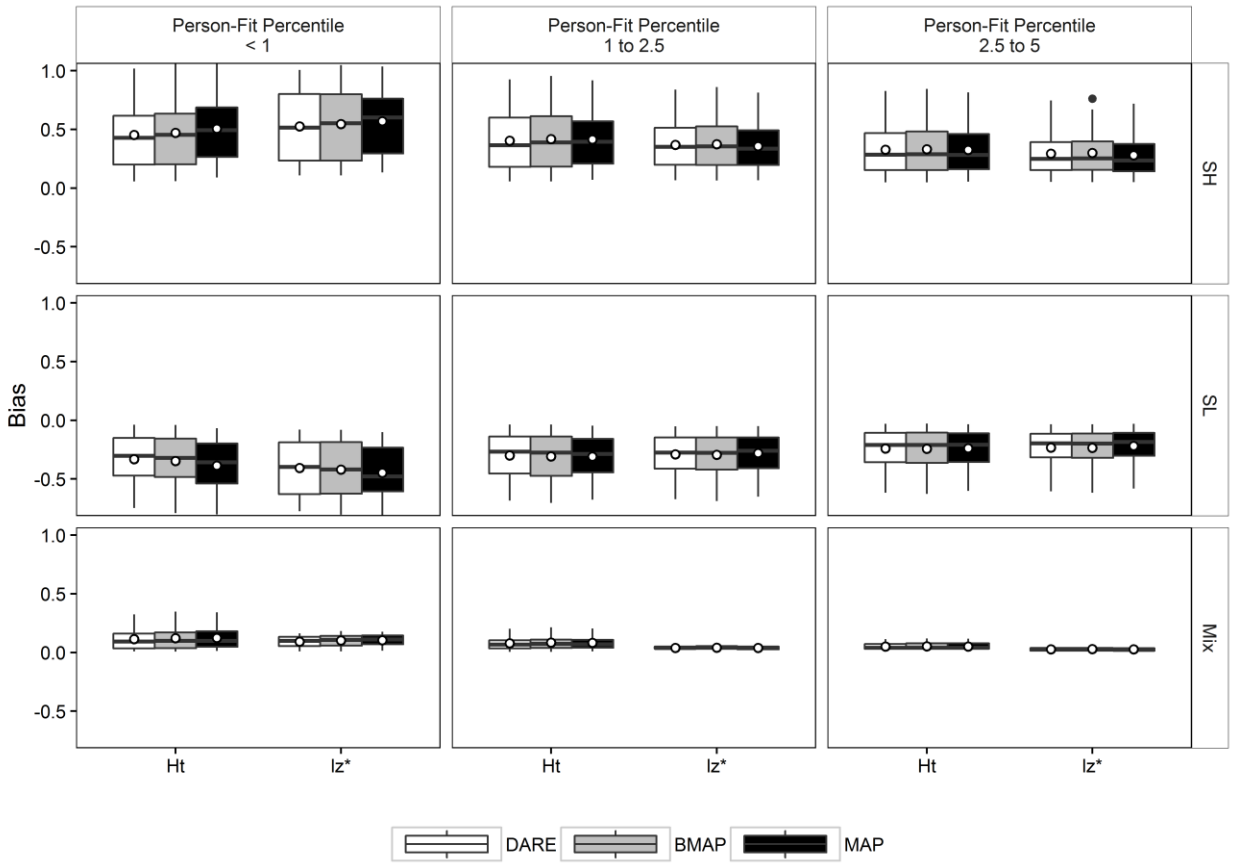

Figure 2. Boxplot of ability estimation bias based on person-fit thresholds and aberrant behavior type

Figure 3 on RMSE tells almost the same story. On average, the DARE and BMAP methods decreased the RMSE over MAP only for examinees meeting the < 1 person-fit percentile criterion. As a result, similar to Maeda and Zhang (2017b), all proceeding results use the .01 α cutoff to show the effects of using robust ability estimation methods among detected examinees.

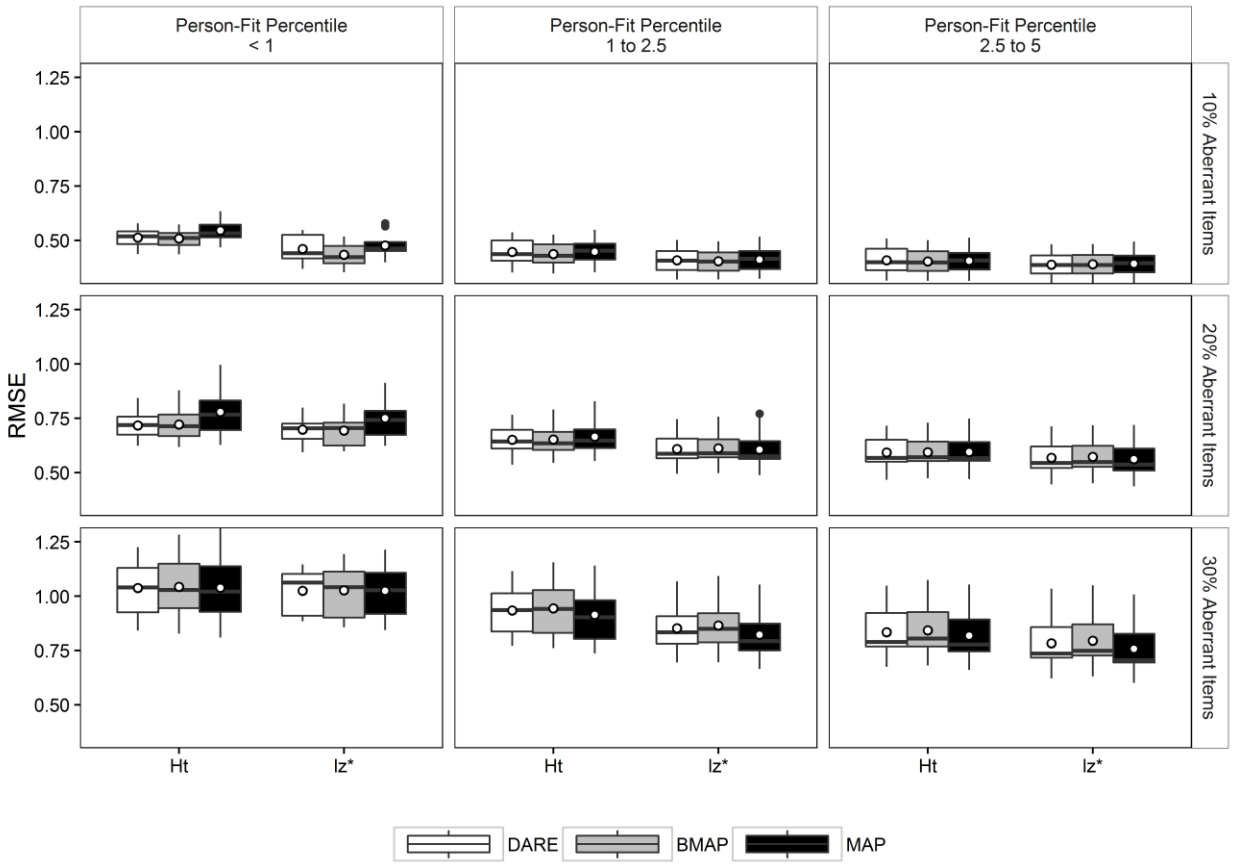

Figure 3. Boxplot of ability estimation RMSE based on person-fit thresholds and percent aberrant items

Ability Estimation Bias

In this section, ability estimation bias are presented by test length. Under each length, results based on both H^T and l_z^* are provided. While l_z^* showed higher power overall, there were situations where H^T outperformed. Besides, showing that l_z^* has more power than H^T does not necessarily prove that l_z^* is superior for the robust estimation purposes. For example, compared to l_z^*, H^T may be better at detecting mildly aberrant examinees who benefit significantly from robust methods. Therefore, an empirical investigation of these issues is useful.

Table 8 shows the bias among examinees detected with H^T under the 20-item test condition. The shaded cell represent the least bias for each condition. When no aberrant

behaviors were present, none of the five estimation methods were biased. When aberrant behaviors were present, the SH conditions showed positive bias, the SL conditions showed negative bias, and the mixed condition showed slight positive bias. These directions were expected given the nature of aberrant behaviors. The biweight method was usually the most biased, followed by MLE. When the percent of aberrant items were 10% or 20%, BMAP and DARE were slightly less biased than MAP overall. This advantage disappeared when the percentage of aberrant items reached 30%. The difference in bias between BMAP and DARE were minute.

Table 8.
Ability Estimation Bias among Examinees Detected with H^T (20 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.00	-0.02	0.01	0.00	0.00
SH	10%	10%	0.14	0.10	0.09	0.06	0.06
		20%	0.36	0.52	0.26	0.23	0.22
		30%	0.61	0.94	0.52	0.53	0.51
	30%	10%	0.24	0.31	0.15	0.11	0.11
		20%	0.46	0.79	0.34	0.30	0.29
		30%	0.71	1.41	0.58	0.62	0.62
SL	10%	10%	-0.12	-0.12	-0.07	-0.04	-0.04
		20%	-0.26	-0.35	-0.19	-0.16	-0.15
		30%	-0.46	-0.68	-0.37	-0.36	-0.35
	30%	10%	-0.20	-0.24	-0.12	-0.08	-0.09
		20%	-0.39	-0.61	-0.27	-0.24	-0.23
		30%	-0.60	-1.09	-0.47	-0.48	-0.48
Mix	10%	10%	0.00	-0.05	0.01	0.01	0.01
		20%	0.04	0.04	0.04	0.03	0.03
		30%	0.13	0.21	0.11	0.12	0.11
	30%	10%	0.04	0.07	0.02	0.01	0.02
		20%	0.07	0.06	0.07	0.06	0.06
		30%	0.19	0.33	0.17	0.17	0.16

Note. AE=aberrant examinees; AI=aberrant items; AB=aberrant behaviors; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

Similar to the H^T results, without aberrant behaviors, none of the estimation methods showed bias when the l_z^* was used in the 20-item test (see Table 9). When aberrant behaviors were present, however, the results deviated from the corresponding H^T conditions. Overall, the bias was slightly larger, which may be because l_z^* detected more misfitting response patterns. In the 10% and 20% aberrant items conditions, the biweight method was often the least biased, MAP usually showed the most bias by a small margin (i.e., 0.0 to 0.02), while MLE, BMAP, and DARE showed about the same degree of bias. In contrast, in the 30% aberrant items conditions, bias was high across the board, indicating that none of the robust methods were effective.

Table 9.
Ability Estimation Bias among Examinees Detected with l_z^* (20 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.00	-0.01	0.00	0.00	-0.01
SH	10%	10%	0.11	0.03	0.13	0.11	0.11
		20%	0.39	0.33	0.40	0.36	0.35
		30%	0.70	0.77	0.70	0.71	0.68
	30%	10%	0.14	0.06	0.18	0.17	0.18
		20%	0.45	0.38	0.49	0.47	0.47
		30%	0.76	0.88	0.78	0.83	0.84
SL	10%	10%	-0.08	-0.03	-0.10	-0.08	-0.08
		20%	-0.29	-0.25	-0.30	-0.27	-0.26
		30%	-0.54	-0.58	-0.55	-0.54	-0.53
	30%	10%	-0.11	-0.03	-0.14	-0.13	-0.14
		20%	-0.35	-0.30	-0.38	-0.37	-0.38
		30%	-0.59	-0.68	-0.62	-0.65	-0.67
Mix	10%	10%	0.01	-0.01	0.01	0.01	0.01
		20%	0.08	0.07	0.08	0.07	0.07
		30%	0.10	0.12	0.10	0.11	0.10
	30%	10%	0.03	0.00	0.04	0.03	0.03
		20%	0.12	0.08	0.12	0.11	0.10
		30%	0.16	0.16	0.16	0.16	0.14

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori;

BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

Compared to the 20-item conditions with H^T, similar estimation biases were observed in the corresponding 40-item test conditions (see Table 10). The key differences were that the biases were larger overall, and the effectiveness of BMAP and DARE improved. In particular, the bias of DARE was the lowest in almost all conditions.

Table 10.
Ability Estimation Bias among Examinees Detected with H^T (40 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	-0.01	-0.02	0.00	0.00	0.00
SH	10%	10%	0.23	0.22	0.17	0.12	0.11
		20%	0.55	0.64	0.47	0.41	0.39
		30%	0.92	1.18	0.83	0.84	0.81
	30%	10%	0.39	0.44	0.27	0.19	0.19
		20%	0.71	0.95	0.57	0.49	0.47
		30%	1.04	1.53	0.90	0.90	0.87
SL	10%	10%	-0.17	-0.17	-0.13	-0.09	-0.08
		20%	-0.40	-0.42	-0.35	-0.28	-0.27
		30%	-0.67	-0.80	-0.62	-0.59	-0.55
	30%	10%	-0.33	-0.36	-0.22	-0.16	-0.16
		20%	-0.56	-0.68	-0.46	-0.39	-0.38
		30%	-0.78	-1.05	-0.70	-0.69	-0.66
Mix	10%	10%	0.03	0.02	0.03	0.02	0.02
		20%	0.11	0.12	0.09	0.08	0.08
		30%	0.22	0.30	0.19	0.21	0.20
	30%	10%	0.07	0.09	0.05	0.04	0.04
		20%	0.17	0.21	0.14	0.13	0.11
		30%	0.33	0.44	0.29	0.30	0.28

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

Comparably, when l_z^* was used with the 40-item conditions, the DARE method showed the smallest bias in many conditions (see Table 11). However, biweight estimation sometimes showed superior performance over DARE, such as in the 10% aberrant examinee / 10% aberrant item conditions, and 30% aberrant examinee / 10% and 20% aberrant item conditions. While the performance of BMAP was comparable to DARE for 10% aberrant items conditions, BMAP was inferior by up to .03 and .05 for the 20% and 30% aberrant items conditions, respectively.

Table 11.
Ability Estimation Bias among Examinees Detected with l_z^* (40 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.00	-0.01	0.00	0.00	-0.01
SH	10%	10%	0.20	0.12	0.22	0.16	0.16
		20%	0.56	0.53	0.57	0.52	0.49
		30%	0.92	0.98	0.91	0.92	0.87
	30%	10%	0.25	0.15	0.28	0.22	0.22
		20%	0.62	0.56	0.64	0.59	0.57
		30%	0.95	0.99	0.95	0.97	0.94
SL	10%	10%	-0.15	-0.09	-0.16	-0.12	-0.12
		20%	-0.44	-0.38	-0.45	-0.39	-0.37
		30%	-0.73	-0.74	-0.73	-0.70	-0.67
	30%	10%	-0.20	-0.12	-0.22	-0.18	-0.18
		20%	-0.49	-0.42	-0.51	-0.46	-0.45
		30%	-0.74	-0.76	-0.75	-0.75	-0.73
Mix	10%	10%	0.03	0.02	0.04	0.03	0.03
		20%	0.10	0.10	0.10	0.10	0.09
		30%	0.13	0.16	0.13	0.14	0.13
	30%	10%	0.07	0.05	0.07	0.06	0.05
		20%	0.15	0.13	0.15	0.14	0.12
		30%	0.18	0.18	0.18	0.18	0.16

Note. AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

When the test length was increased to 60 items, using H^T, DARE continued to show the least bias in almost all conditions (see Table 12). The performance of DARE was often followed

by BMAP, MAP, MLE, then biweight, in that order. The exceptions to these were the 50% aberrant items conditions, where MAP showed the lowest bias. As expected, widening the spread of item difficulty (i.e., variance of 4) worked favorably for all three robust estimation methods.

Table 12.
Ability Estimation Bias among Examinees Detected with H^T (60 Items)

σ^2	AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
1	None	0%	0%	0.00	0.00	0.00	0.00	0.00
	SH	10%	10%	0.28	0.25	0.23	0.17	0.16
			20%	0.66	0.69	0.60	0.53	0.49
			30%	1.05	1.25	0.99	1.00	0.95
		30%	10%	0.46	0.46	0.36	0.26	0.26
			20%	0.81	0.94	0.72	0.64	0.60
			30%	1.18	1.52	1.06	1.06	1.02
	SL	10%	10%	-0.21	-0.18	-0.17	-0.12	-0.11
			20%	-0.48	-0.47	-0.44	-0.37	-0.34
			30%	-0.77	-0.85	-0.74	-0.72	-0.66
		30%	10%	-0.37	-0.37	-0.29	-0.21	-0.21
			20%	-0.62	-0.67	-0.56	-0.48	-0.46
			30%	-0.85	-1.00	-0.80	-0.79	-0.75
	Mix	10%	10%	0.06	0.06	0.05	0.03	0.03
			20%	0.16	0.17	0.14	0.14	0.12
			30%	0.25	0.32	0.23	0.25	0.24
			50%	0.39	0.57	0.35	0.40	0.42
		30%	10%	0.08	0.07	0.08	0.06	0.05
			20%	0.21	0.24	0.18	0.17	0.15
			30%	0.38	0.48	0.34	0.35	0.32
			50%	0.71	0.96	0.64	0.68	0.71
4	Mix	30%	10%	0.11	0.07	0.11	0.07	0.07
			20%	0.26	0.25	0.25	0.23	0.19
			30%	0.42	0.48	0.39	0.41	0.38

Note. σ^2=variance of the item difficulty parameter; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

In the 60-item conditions using l_z^* to detect aberrant examinees, DARE was almost always associated with the lowest bias. The only exceptions were the 10% aberrant items conditions where the biweight method outperformed (see Table 13).

Table 13.
Ability Estimation Bias among Examinees Detected with l_z^* (60 Items)

σ^2	AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
1	None	0%	0%	0.00	0.00	0.00	0.00	0.00
	SH	10%	10%	0.25	0.17	0.26	0.20	0.19
			20%	0.64	0.60	0.64	0.59	0.54
			30%	1.00	1.07	0.99	1.01	0.96
		30%	10%	0.32	0.22	0.34	0.27	0.26
			20%	0.69	0.64	0.70	0.65	0.62
			30%	1.03	1.06	1.04	1.05	1.01
	SL	10%	10%	-0.19	-0.12	-0.20	-0.15	-0.14
			20%	-0.51	-0.45	-0.51	-0.45	-0.41
			30%	-0.81	-0.82	-0.81	-0.79	-0.73
		30%	10%	-0.26	-0.17	-0.27	-0.21	-0.21
			20%	-0.56	-0.49	-0.57	-0.51	-0.49
			30%	-0.81	-0.82	-0.82	-0.80	-0.78
	Mix	10%	10%	0.06	0.04	0.06	0.04	0.04
			20%	0.12	0.13	0.12	0.12	0.11
			30%	0.12	0.15	0.12	0.14	0.13
			50%	-0.01	0.04	-0.01	0.02	0.04
		30%	10%	0.09	0.06	0.09	0.07	0.06
			20%	0.16	0.15	0.16	0.15	0.13
			30%	0.16	0.17	0.16	0.17	0.15
			50%	-0.08	-0.08	-0.08	-0.07	-0.06
4	Mix	30%	10%	0.10	0.07	0.11	0.07	0.07
			20%	0.17	0.15	0.17	0.15	0.13
			30%	0.10	0.10	0.10	0.11	0.09

Note. σ^2=variance of the item difficulty parameter; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest absolute bias values for each row are highlighted.

Overall, across all conditions, all robust methods showed less bias than MAP when the test length was increased. DARE was often associated with the lowest bias, followed closely by BMAP. The biweight method showed the most obscure results in that its bias was favorable when used with l_z^* but not with H^T. The decrease in bias by using DARE instead of MAP was about equal for l_z^* and H^T.

Ability Estimation RMSE

For the 20-item test with H^T, as expected, the RMSE of the ability estimates increased quickly with more aberrant items (see Table 14). Overall, MLE and biweight estimates consistently had much higher RMSE than MAP, BMAP, and DARE. In particular, the RMSE of the biweight method was always intolerably high (i.e., much higher than 1), frequently reaching over 2.0. When no aberrant behaviors were present, MAP had the lowest RMSE at 0.47, while BMAP and DARE had slightly higher RMSE at 0.50 and 0.51, respectively. MAP also had the lowest RMSE for most 10% aberrant examinee conditions by a small margin, while BMAP had the lowest RMSE for most 30% aberrant examinee conditions. Across most conditions, the RMSE of DARE was the highest among these three methods, but the differences were negligible.

Table 14.
Ability Estimation RMSE among Examinees Detected with H^T (20 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.78	1.75	0.47	0.50	0.51
SH	10%	10%	0.80	1.80	0.53	0.53	0.54
		20%	0.93	1.90	0.69	0.68	0.69
		30%	1.14	2.14	0.93	0.97	0.98
	30%	10%	0.79	1.85	0.57	0.55	0.56
		20%	0.92	2.01	0.73	0.70	0.72
		30%	1.13	2.35	0.93	0.97	1.01
SL	10%	10%	0.79	1.77	0.52	0.53	0.53
		20%	0.87	1.84	0.63	0.62	0.63
		30%	1.03	1.99	0.81	0.83	0.84
	30%	10%	0.80	1.83	0.56	0.55	0.56
		20%	0.90	1.94	0.68	0.65	0.67
		30%	1.07	2.16	0.84	0.86	0.89
Mix	10%	10%	0.81	1.80	0.53	0.54	0.54
		20%	0.94	1.88	0.69	0.67	0.68
		30%	1.16	2.10	0.94	0.95	0.95
	30%	10%	0.84	1.85	0.62	0.57	0.58
		20%	1.04	2.05	0.83	0.77	0.77
		30%	1.31	2.39	1.10	1.09	1.08

Note. RMSE=root-mean-squared-error; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest RMSE values for each row are highlighted.

Among examinees detected with l_z^* for the 20-item test, the RMSE of MLE and biweight was much lower than in the corresponding H^T conditions (see Table 15). However, the RMSE of the biweight method was still too high to be useful. MLE, MAP, and BMAP often had the lowest RMSE in most conditions, and the differences between the three were typically very small. The RMSE for DARE was overall slightly higher than these three methods. Otherwise, the patterns of RMSE were similar to the corresponding H^T conditions.

Table 15.
Ability Estimation RMSE among Examinees Detected with l_z^* (20 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.40	0.98	0.37	0.45	0.51
SH	10%	10%	0.47	1.01	0.47	0.48	0.53
		20%	0.71	1.10	0.71	0.68	0.71
		30%	1.01	1.41	0.99	1.01	1.03
	30%	10%	0.47	1.07	0.50	0.50	0.55
		20%	0.69	1.04	0.73	0.70	0.75
		30%	0.96	1.31	0.98	1.02	1.07
SL	10%	10%	0.45	1.00	0.45	0.47	0.53
		20%	0.62	1.06	0.62	0.61	0.64
		30%	0.85	1.27	0.84	0.86	0.88
	30%	10%	0.46	1.06	0.48	0.48	0.53
		20%	0.66	1.09	0.67	0.64	0.69
		30%	0.83	1.26	0.85	0.88	0.93
Mix	10%	10%	0.48	1.02	0.48	0.47	0.52
		20%	0.73	1.11	0.72	0.68	0.70
		30%	1.03	1.43	1.00	0.99	1.00
	30%	10%	0.55	1.03	0.58	0.52	0.55
		20%	0.87	1.15	0.88	0.80	0.80
		30%	1.19	1.50	1.18	1.15	1.13

Note. RMSE=root-mean-squared-error; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest RMSE values for each row are highlighted.

Among examinees detected with H^T, when the test length was increased from 20 to 40 items, BMAP and DARE improved effectiveness in decreasing the RMSE relative to MAP (see Table 16). BMAP and DARE equally and consistently showed lower RMSE than MAP in the 10% and 20% aberrant items conditions. MAP had the superior RMSE when aberrant behavior was not present in the data. MAP, BMAP, and DARE were about equally accurate at 30% aberrant items conditions. The exception was that compared to MAP, BMAP and DARE had lower RMSE by 0.04 and 0.07, respectively, in the 30% aberrant examinee, 30% aberrant item, and mixed type conditions. Similar to the 20-item conditions, MLE and biweight methods were still inaccurate.

Table 16.
Ability Estimation RMSE among Examinees Detected with H^T (40 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.65	1.20	0.38	0.43	0.43
SH	10%	10%	0.73	1.24	0.51	0.49	0.49
		20%	0.95	1.42	0.77	0.73	0.73
		30%	1.26	1.82	1.11	1.13	1.13
	30%	10%	0.78	1.32	0.57	0.51	0.52
		20%	1.01	1.60	0.83	0.76	0.76
		30%	1.28	2.05	1.10	1.10	1.11
SL	10%	10%	0.70	1.22	0.48	0.47	0.47
		20%	0.84	1.29	0.67	0.62	0.62
		30%	1.04	1.49	0.92	0.91	0.90
	30%	10%	0.75	1.30	0.54	0.50	0.50
		20%	0.89	1.41	0.73	0.67	0.67
		30%	1.05	1.62	0.92	0.91	0.92
Mix	10%	10%	0.73	1.23	0.51	0.49	0.49
		20%	0.96	1.39	0.79	0.73	0.72
		30%	1.23	1.71	1.08	1.09	1.07
	30%	10%	0.82	1.31	0.62	0.54	0.55
		20%	1.10	1.55	0.95	0.84	0.83
		30%	1.42	1.99	1.27	1.23	1.20

Note. RMSE=root-mean-squared-error; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest RMSE values for each row are highlighted

Much like in the H^T 40-item conditions, when l_z^* was used to detect aberrant examinees, BMAP and DARE equally and consistently showed lower RMSE than MAP in the 10% and 20% aberrant items conditions. However, MLE and biweight showed much higher estimation accuracy in these conditions than when H^T was used. The RMSE for MLE was nearly the same as that of MAP in all conditions. Biweight still showed the highest RMSE out of all the methods studied.

Table 17.
Ability Estimation RMSE among Examinees Detected with l_z^* (40 Items)

AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
None	0%	0%	0.28	0.59	0.27	0.32	0.37
SH	10%	10%	0.44	0.62	0.45	0.41	0.43
		20%	0.75	0.87	0.76	0.71	0.71
		30%	1.10	1.29	1.09	1.11	1.10
	30%	10%	0.45	0.64	0.48	0.43	0.45
		20%	0.75	0.80	0.77	0.71	0.72
		30%	1.05	1.16	1.06	1.07	1.08
SL	10%	10%	0.40	0.62	0.41	0.38	0.41
		20%	0.65	0.76	0.65	0.60	0.61
		30%	0.92	1.06	0.91	0.89	0.89
	30%	10%	0.42	0.66	0.45	0.41	0.43
		20%	0.64	0.75	0.67	0.61	0.63
		30%	0.86	1.00	0.87	0.87	0.89
Mix	10%	10%	0.44	0.63	0.46	0.41	0.43
		20%	0.78	0.86	0.78	0.71	0.70
		30%	1.10	1.28	1.08	1.08	1.05
	30%	10%	0.53	0.65	0.57	0.47	0.47
		20%	0.90	0.91	0.91	0.81	0.78
		30%	1.22	1.32	1.21	1.19	1.15

Note. RMSE=root-mean-squared-error; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; The lowest RMSE values for each row are highlighted.

Increasing the test length from 40 to 60 items in the H^T conditions decreased the overall RMSE for all methods, but much of the relative patterns still remained the same (see Table 18). One notable change was that DARE was now associated with the lowest RMSE in almost all aberrant behavior conditions. BMAP was the next highest in accuracy, while MAP was the third. An exception to this was in the 10% aberrant examinee 30% aberrant items conditions, where BMAP showed higher RMSE than MAP. RMSE in the 50% aberrant items conditions for all methods were excessively high. In the conditions where the item difficulty parameters were

sampled from a $N(0,4)$ distribution, both BMAP and DARE had much lower RMSE than MAP by up to 0.19.

Table 18.
Ability Estimation RMSE among Examinees Detected with H^T (60 Items)

σ^2	AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
1	None	0%	0%	0.52	0.93	0.34	0.38	0.38
	SH	10%	10%	0.67	0.98	0.51	0.46	0.47
			20%	0.96	1.22	0.83	0.78	0.76
			30%	1.29	1.66	1.19	1.21	1.19
		30%	10%	0.76	1.07	0.59	0.51	0.52
			20%	1.04	1.37	0.89	0.81	0.79
			30%	1.35	1.87	1.20	1.19	1.18
	SL	10%	10%	0.62	0.95	0.47	0.44	0.44
			20%	0.79	1.04	0.70	0.64	0.62
			30%	1.02	1.30	0.96	0.95	0.92
		30%	10%	0.68	1.03	0.53	0.48	0.48
			20%	0.85	1.14	0.76	0.68	0.68
			30%	1.01	1.34	0.95	0.94	0.93
	Mix	10%	10%	0.67	0.97	0.51	0.47	0.46
			20%	0.93	1.16	0.82	0.75	0.73
			30%	1.24	1.55	1.15	1.15	1.12
			50%	1.95	2.59	1.85	2.06	2.10
		30%	10%	0.77	1.03	0.63	0.53	0.53
			20%	1.11	1.33	1.00	0.88	0.84
			30%	1.41	1.75	1.32	1.28	1.23
			50%	2.08	2.68	1.97	2.10	2.16
4	Mix	30%	10%	0.64	0.62	0.63	0.48	0.48
			20%	1.09	1.06	1.05	0.92	0.86
			30%	1.53	1.67	1.47	1.44	1.37

Note. RMSE=root-mean-squared-error; σ^2=variance of the item difficulty parameter; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest RMSE values for each row are highlighted.

In the l_z^* conditions with the 60-item tests, DARE typically showed the lowest RMSE in all 20% and 30% aberrant items conditions (see Table 19). In these conditions, the difference in RMSE between DARE and MAP was greatest in the mixed behavior condition. BMAP was

often associated with the lowest RMSE in most 10% aberrant items conditions, but DARE was close behind (i.e., up to a 0.02 difference). MAP and MLE had the lowest RMSE in the 0% and 50% aberrant items conditions. Although there were some exceptions, biweight was typically the least accurate method.

Table 19.
Ability Estimation RMSE among Examinees Detected with l_z^* (60 Items)

σ^2	AB	AE	AI	MLE	Biweight	MAP	BMAP	DARE
1	None	0%	0%	0.24	0.45	0.23	0.27	0.31
	SH	10%	10%	0.44	0.50	0.45	0.39	0.40
			20%	0.78	0.82	0.78	0.73	0.71
			30%	1.14	1.27	1.13	1.15	1.13
		30%	10%	0.47	0.51	0.49	0.42	0.43
			20%	0.79	0.77	0.80	0.74	0.73
			30%	1.11	1.16	1.11	1.11	1.10
	SL	10%	10%	0.38	0.49	0.40	0.35	0.37
			20%	0.66	0.69	0.67	0.60	0.59
			30%	0.94	1.04	0.94	0.92	0.90
		30%	10%	0.42	0.52	0.44	0.38	0.40
			20%	0.67	0.68	0.69	0.62	0.62
			30%	0.89	0.95	0.90	0.89	0.89
	Mix	10%	10%	0.44	0.50	0.45	0.39	0.39
			20%	0.79	0.81	0.78	0.72	0.69
			30%	1.12	1.26	1.10	1.11	1.07
			50%	1.81	2.28	1.77	1.98	2.00
		30%	10%	0.54	0.52	0.56	0.45	0.45
			20%	0.91	0.86	0.91	0.82	0.78
			30%	1.22	1.28	1.21	1.19	1.14
			50%	1.87	2.13	1.84	1.97	2.01
4	Mix	30%	10%	0.58	0.49	0.61	0.47	0.46
			20%	1.02	0.94	1.01	0.89	0.83
			30%	1.41	1.46	1.39	1.38	1.31

Note. RMSE=root-mean-squared-error; σ^2=variance of the item difficulty parameter; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high; MLE=maximum likelihood estimation; MAP=maximum a-posteriori; BMAP=biweight-MAP; DARE=downweighting of aberrant responses estimation; the lowest RMSE values for each row are highlighted.

Overall, assessment of estimation accuracy across all conditions and methods were similar between RMSE and bias; high RMSE was often associated with high bias, and vice versa. The rankings of the estimation accuracy among the five studied methods for a given condition were consistently similar between RMSE and bias, with the exception of biweight having the lowest bias in many conditions while having the highest RMSE in almost all conditions. All robust methods showed improved RMSE relative to MAP when the test length increased. In the 10% to 30% aberrant items conditions, DARE was the most frequently associated with the lowest RMSE, followed closely by BMAP. The biweight method was largely ineffective in decreasing RMSE relative to MAP in almost all conditions. Similar to bias, the decrease in RMSE by using DARE instead of MAP was about equal for l_z^* and H^T. Therefore, given that l_z^* had consistently higher power and lower Type I error than H^T, DARE may synergize better with l_z^*. By using l_z^* rather than H^T, DARE can lower the ability estimation bias and RMSE among more examinees with aberrant behavior while avoiding decreases in estimation accuracy among the non-aberrant examinees. Therefore, all proceeding analyses were focused on using the l_z^* with DARE.

Ability Estimation Bias by Ability Levels

In this section, ability estimation bias is presented by ability levels for limited testing conditions. The goal was to determine whether the effects of robust estimation on bias observed in the previous sections applied similarly to all ability levels. Since high bias and/or RMSE was observed for MLE and biweight, they were excluded. Also excluded were the 0% and 50% aberrant items conditions where robust estimation methods were ineffective. The 60-item SH and mixed aberrant behavior conditions were selected for illustration purpose. Under these conditions, DARE performed fairly well.

Examinees were grouped first by the true θ then by the MAP estimate of θ. Grouping by true θ can reveal differential effects of the robust methods among examinees at various trait levels. In the situations where this is the case, further actions may be necessary to obtain accurate estimates for every ability level. Unfortunately, the true θ will not be always helpful as it is generally unknown in practice. One obvious substitute is the MAP estimate. For example, if DARE is shown to be effective only among examinees with MAP of near 0, practitioners can actually use this finding to improve ability estimation for those examinees.

Figure 4 shows the estimation bias by the ability level using the true θ for the SH conditions. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the detected true aberrant examinees. It seems that BMAP and DARE had improved bias compared to MAP only among examinees with low θ. For these examinees, DARE was slightly less biased than BMAP in the 20% and 30% aberrant items conditions. MAP was superior to BMAP and DARE for examinees with θ of 0 or more. These seemingly disappointing results were actually not disappointing at all. Due to the nature of aberrance, most detected examinees (i.e., usually over 70%), especially those that were actually aberrant, had low θ. This pattern was more pronounced as the percentage of aberrant examinees and items increased. For example, the 30% aberrant examinee 20% aberrant item conditions in Figure 4 show that only 16% of the detected examinees had θ of 0 or more. Therefore, DARE effectively reduced estimation bias compared to BMAP and MAP in the most essential levels of θ. An additional observation is that all estimation methods showed increased bias at the low levels of θ. This was expected because examinees need incorrect responses for them to spuriously increase test scores.

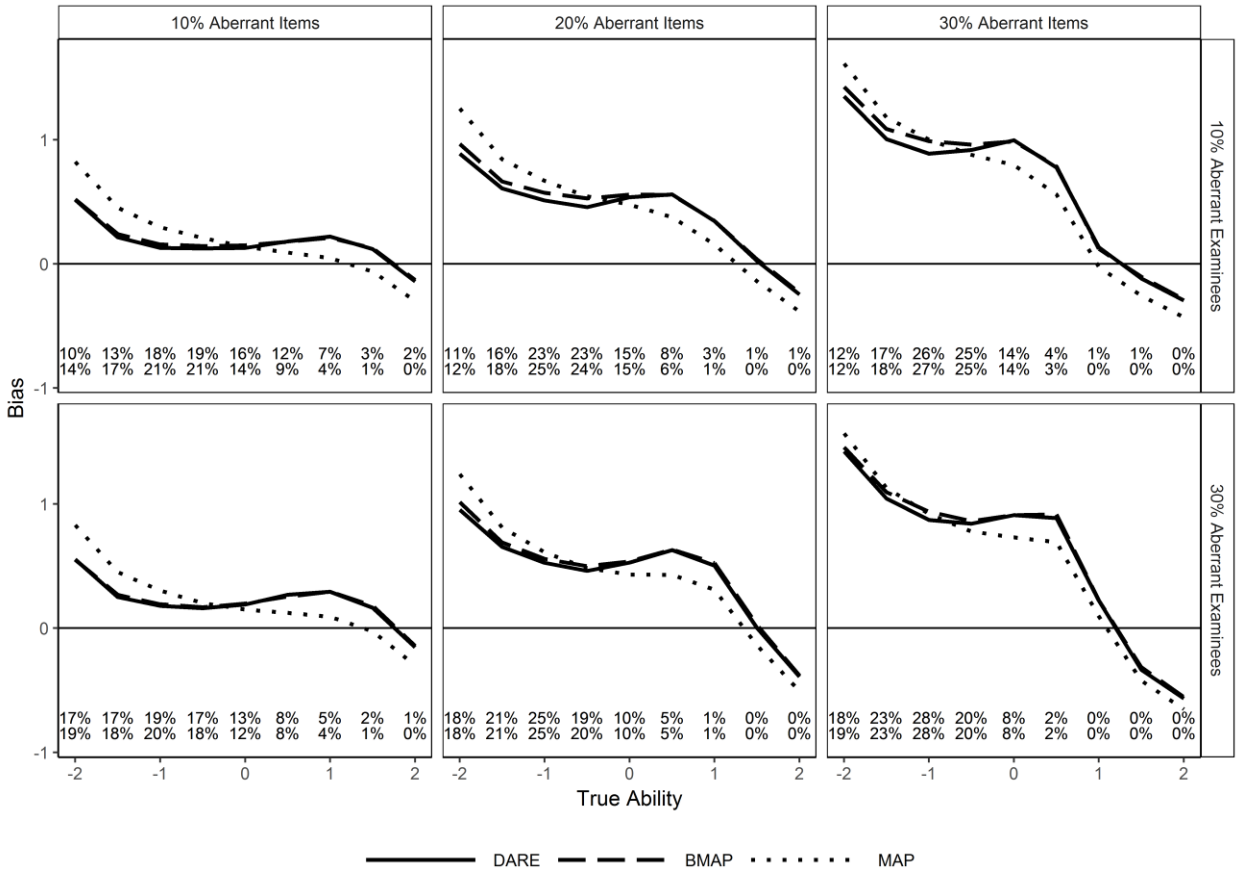

Figure 4. Ability estimation bias by θ (60 items, SH behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Figure 5 presents the same results as Figure 4, but using the MAP estimates. The majority of examinees detected in the SH behavior conditions had an ability estimate around -1.0 to 0.0. These examinees benefited the most from DARE. In some infrequent cases, when aberrant correct responses spuriously increased the examinee MAP to about 0.0 or more, BMAP and DARE were unable to revise these response patterns. The combination of Figures 4 and 5 showed that DARE was most effective for examinees with both θ and spuriously increased MAP estimate of less than 0.0.

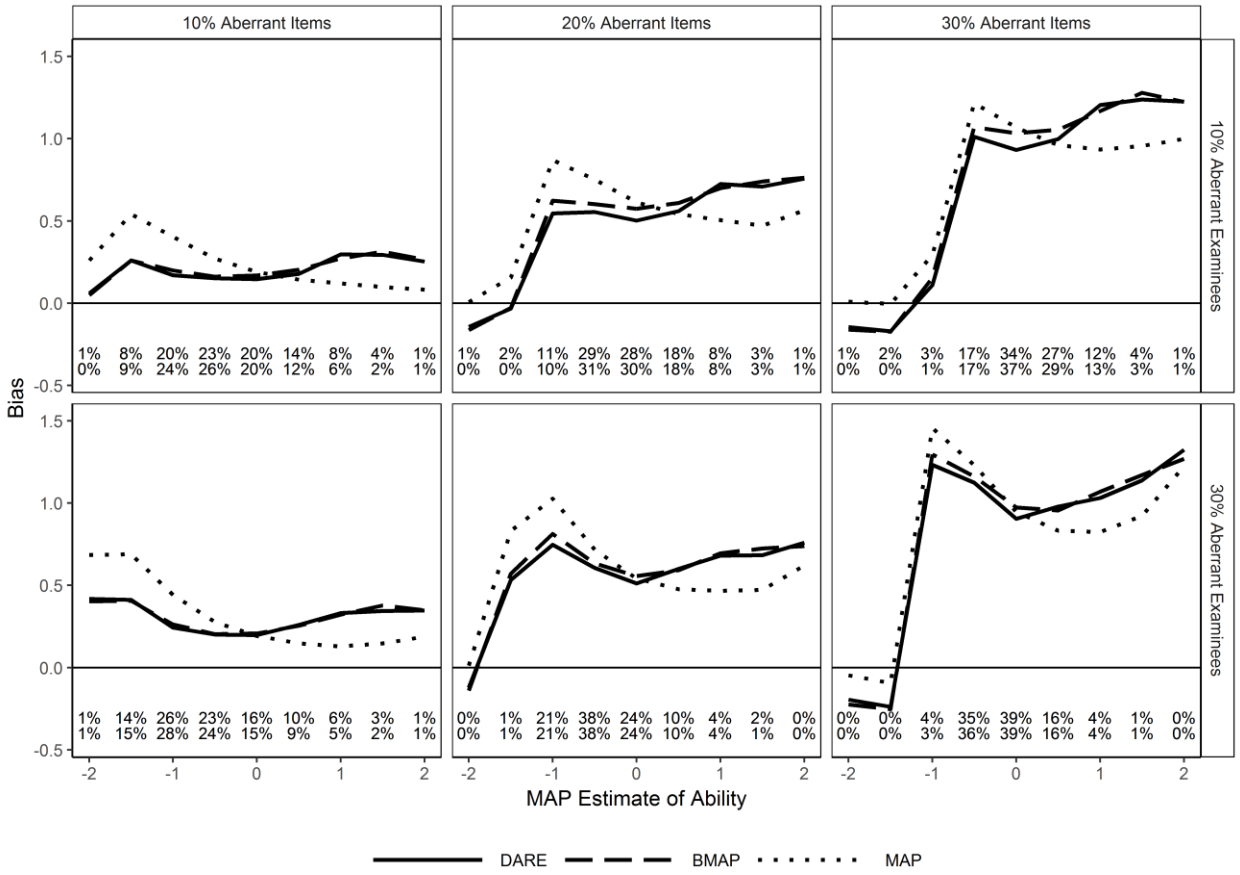

Figure 5. Ability estimation bias by MAP ability (60 items, SH behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

In the 60-item mixed aberrant behavior conditions, all estimation methods continued to show increased bias at the extreme levels of θ (see Figure 6). BMAP and DARE showed less bias than MAP at all levels of θ, especially in the extreme ends. DARE was less biased than BMAP in the 20% and 30% aberrant items conditions. The distribution of detected examinees as well as the distribution of detected aberrant examinees was spread out across all θ levels.

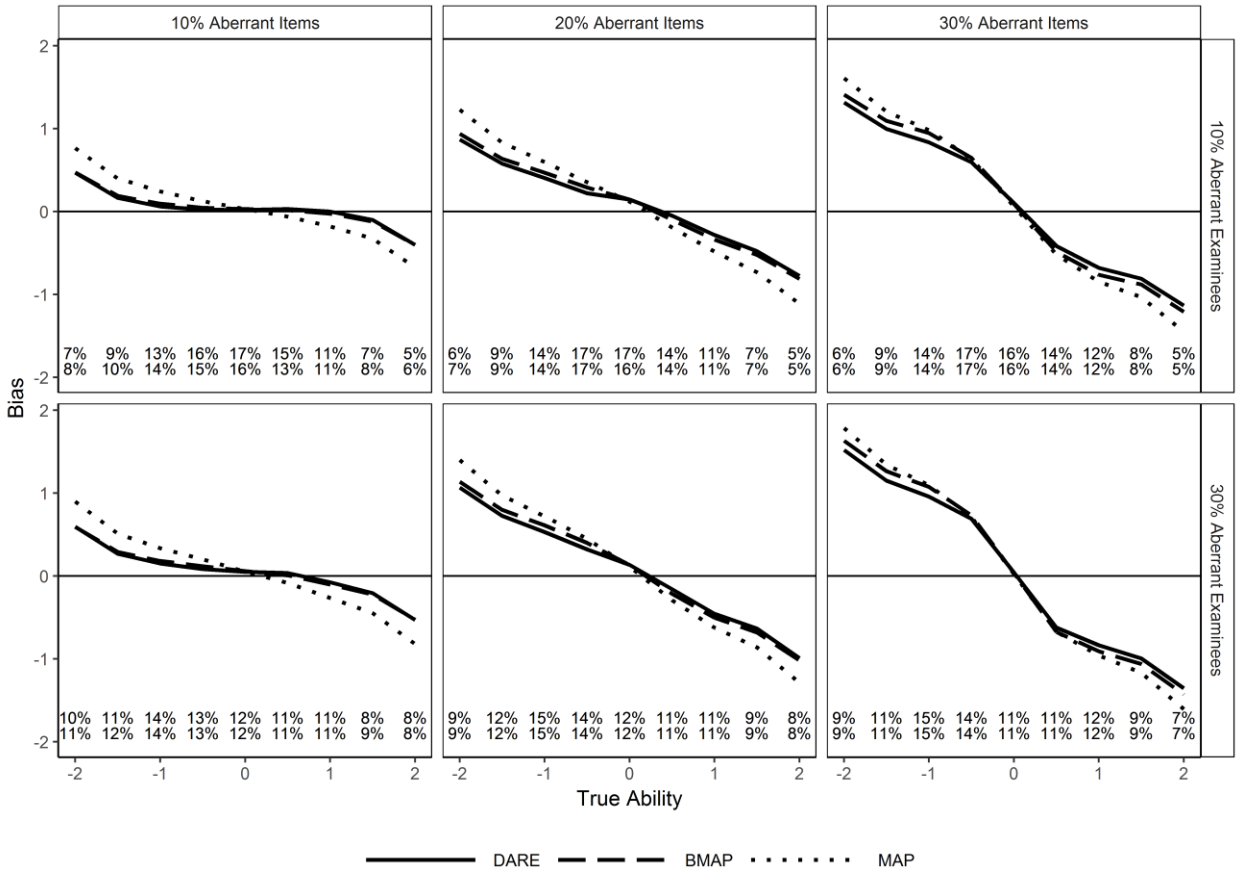

Figure 6. Ability estimation bias by θ (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Examining the same conditions by MAP instead of θ revealed a very different pattern (see Figure 7). For example, the distribution of detected examinees tended to gather in the center of the distribution (i.e., -0.5 to 0.5), especially in the 30% aberrant items conditions. Interestingly, BMAP and DARE were more biased than MAP in some situations, such as 30% aberrant examinee 20% aberrant item condition at MAP of -2.0, -1.5, 1.5, and 2.0. However, only a handful of people were in those MAP categories (i.e., 4% total), which showed that DARE and BMAP were effective in reducing bias for the majority of the examinees.

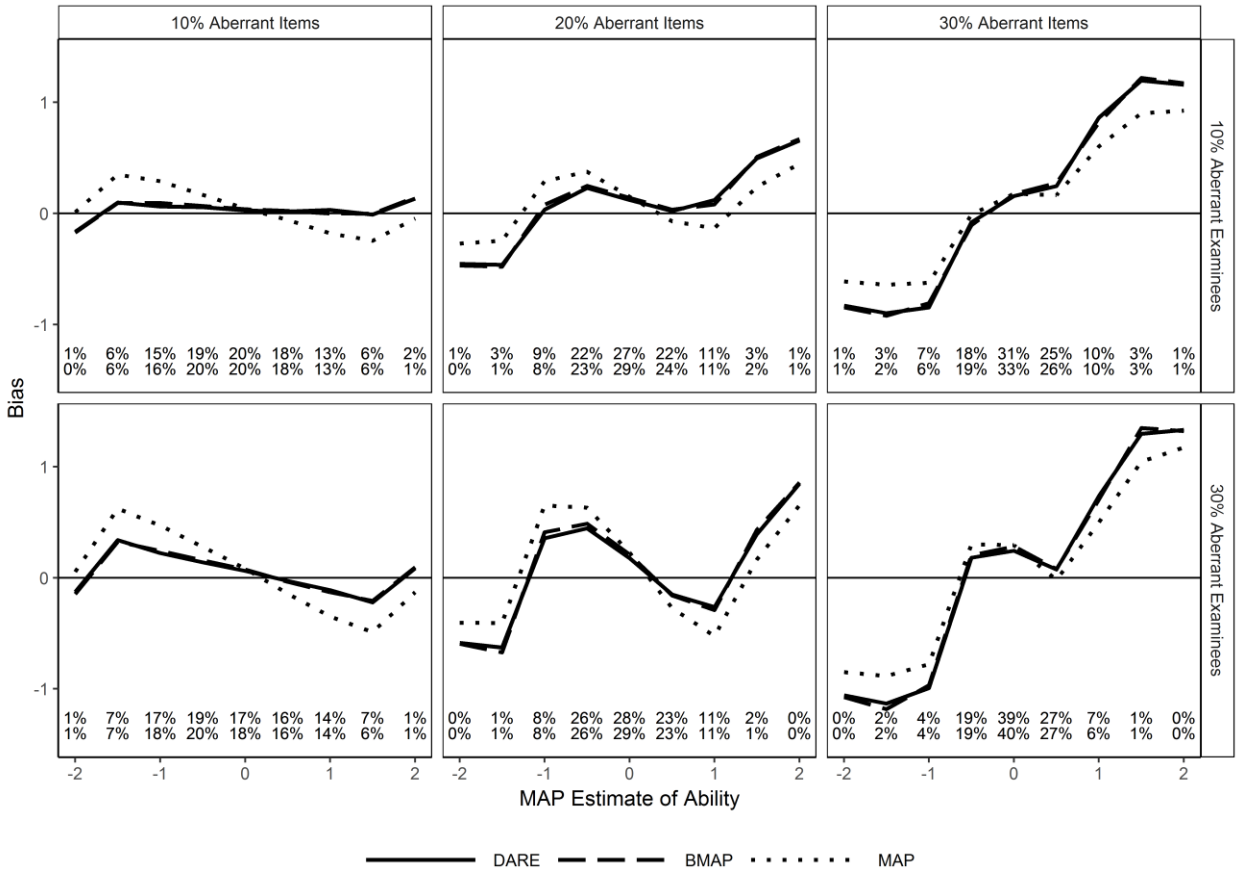

Figure 7. Ability estimation bias by MAP ability (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Bias by θ in the mixed aberrant behavior conditions with item difficulty parameters sampled from a $N(0,4)$ distribution showed similar patterns to those observed in the corresponding conditions with a $N(0,1)$ distribution. However, wider (see Figure 8) compared to narrower (see Figure 6) spread of item difficulty was associated with improved effectiveness of both BMAP and DARE. These improvements were pronounced among examinees with extreme θ. Finally, DARE was clearly more effective than BMAP in the 20% and 30% aberrant items conditions.

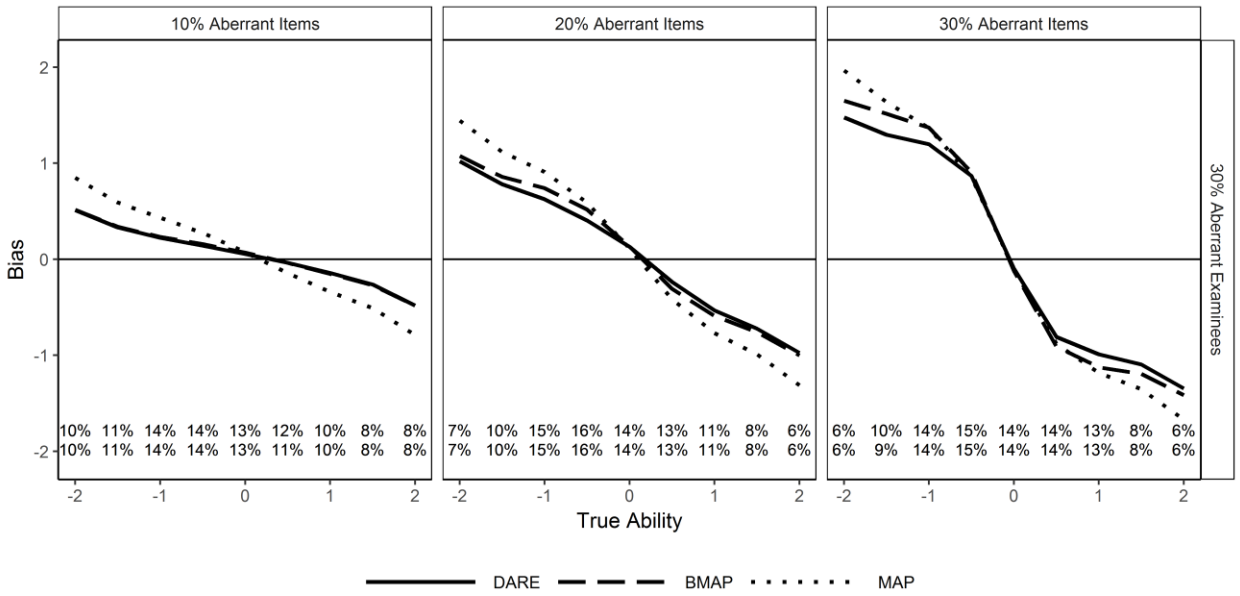

Figure 8. Ability estimation bias by θ with item difficulty variance of 4 (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

The pattern of bias by MAP with item difficulty parameters sampled from a $N(0,4)$ distribution showed interesting results (see Figure 9). BMAP and DARE were effective in decreasing bias compared to MAP in the 10% aberrant items condition. In the 20% and 30% aberrant items condition, DARE had the lowest bias in the center of the distribution where most people belonged (i.e., MAP of -0.5 to 0.5). However, with 30% aberrant items, BMAP was never as good as MAP regardless of the examinee MAP level. These results were inconsistent with Figure 8 and Table 12 under the same conditions, where BMAP showed bias that was at least as low as MAP. Perhaps the positive and negative biases of examinees were canceling each other to show low bias from limited perspectives. Similar to previous results, BMAP and DARE had worse bias than MAP at the extreme levels of MAP.

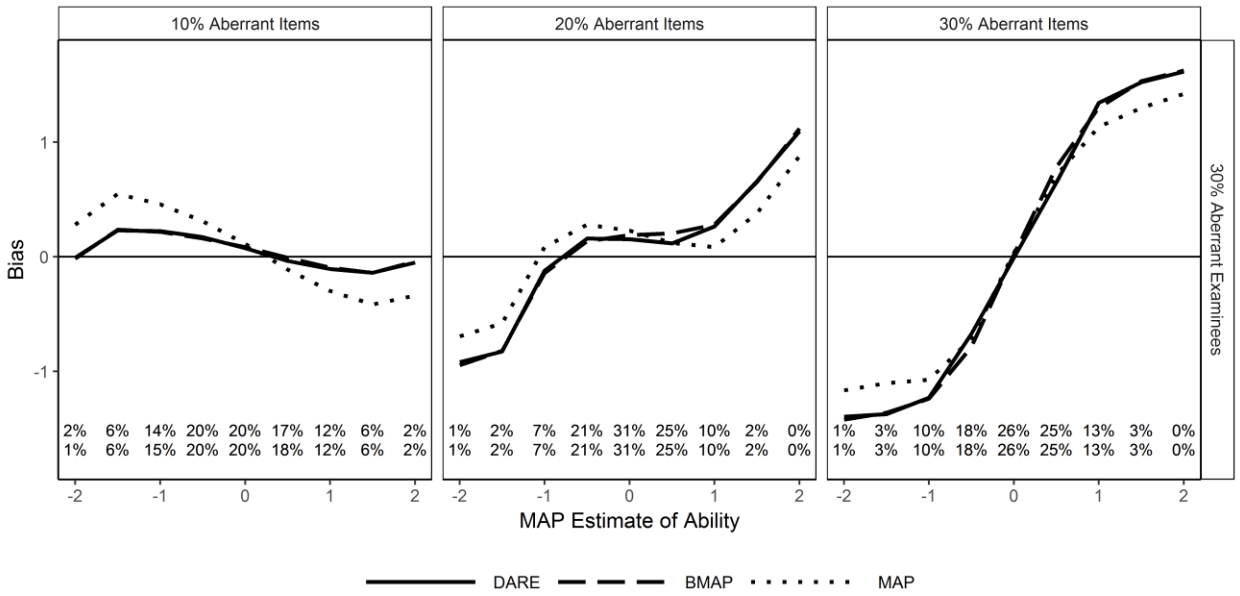

Figure 9. Ability estimation bias by MAP ability with item difficulty variance of 4 (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Ability Estimation RMSE by Ability Levels

Similar to the previous section, the ability estimation RMSE of MAP, BMAP, and DARE among examinees detected with l_z^* were further examined by true and MAP ability levels. In the 60-item SH behavior conditions, RMSE was generally higher for examinees with lower θ. These examinees benefited most from BMAP and DARE compared to MAP, while MAP was superior among detected examinees with θ of about 0.0 or more.

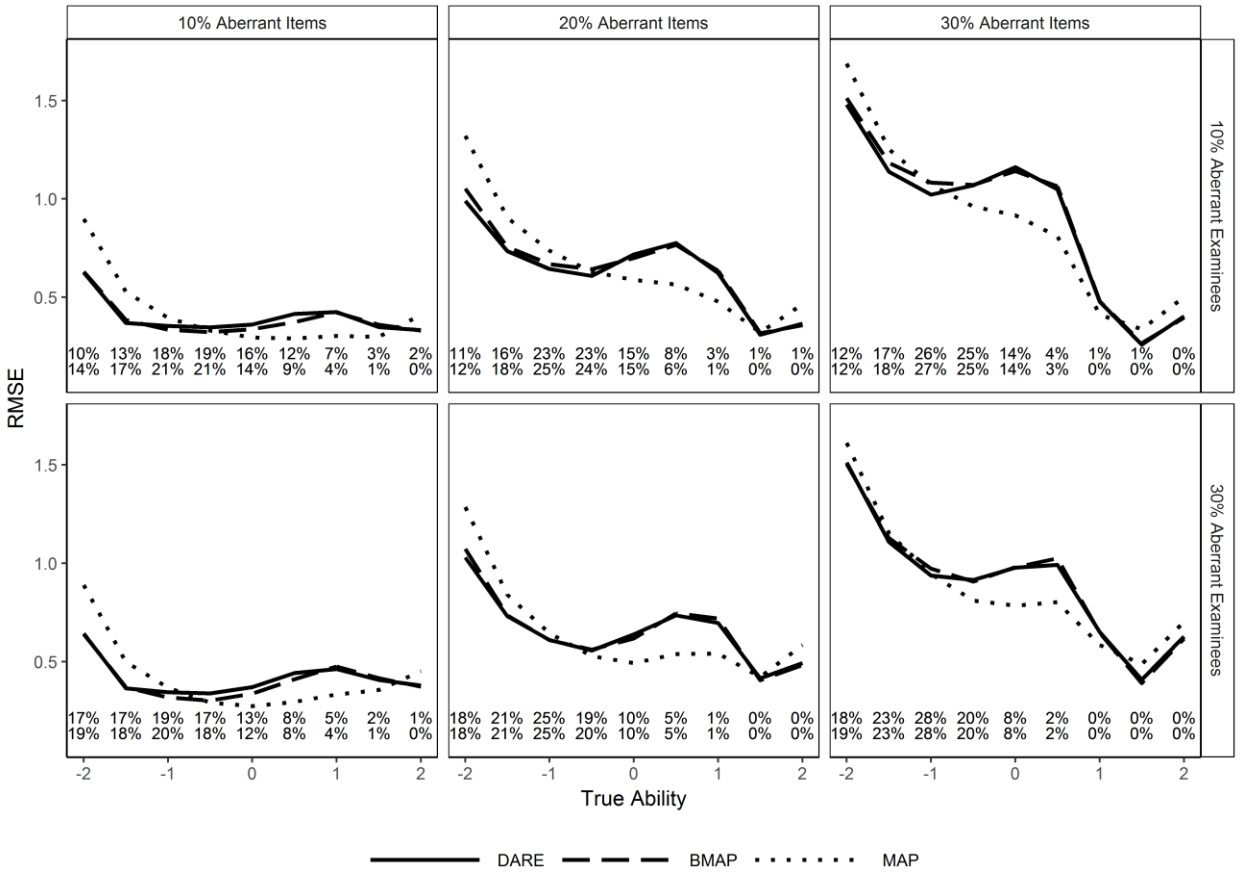

Figure 10. Ability estimation RMSE by θ (60 items, SH behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Examining the same data by levels of MAP showed a similar pattern where BMAP and DARE was associated with lower RMSE than MAP among examinees with MAP of about 0.0 or less, which was typically no less than 60% of the detected examinees (see Figure 11). MAP had the lowest RMSE among examinees with MAP of over 0.0. BMAP usually had slightly lower RMSE than DARE in the 10% aberrant items conditions, while DARE was slightly more effective in the 20% and 30% aberrant items conditions.

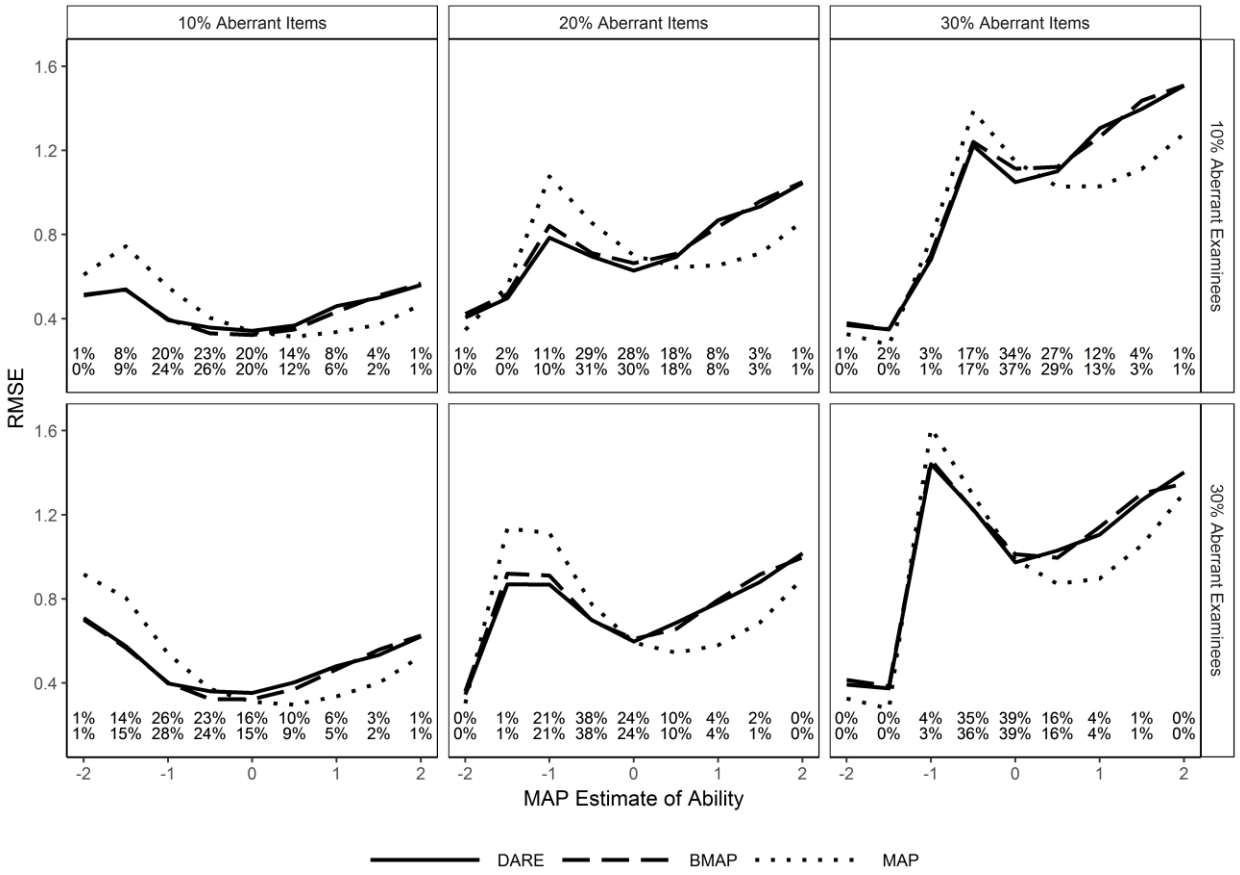

Figure 11. Ability estimation RMSE by MAP ability (60 items, SH behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

RMSE by θ in the mixed aberrant behavior conditions showed patterns completely different from that of the SH conditions (see Figure 12). In all of these conditions, using BMAP or DARE instead of MAP was effective in reducing RMSE among examinees with θ of under -0.5 or over 0.5. This effectiveness increased as the θ became more extreme, which was important because a considerable proportion (e.g., 30%) of examinees had extreme θ values. Finally, DARE had lower RMSE than BMAP in the 20% and 30% aberrant items conditions across the entire range of θ.

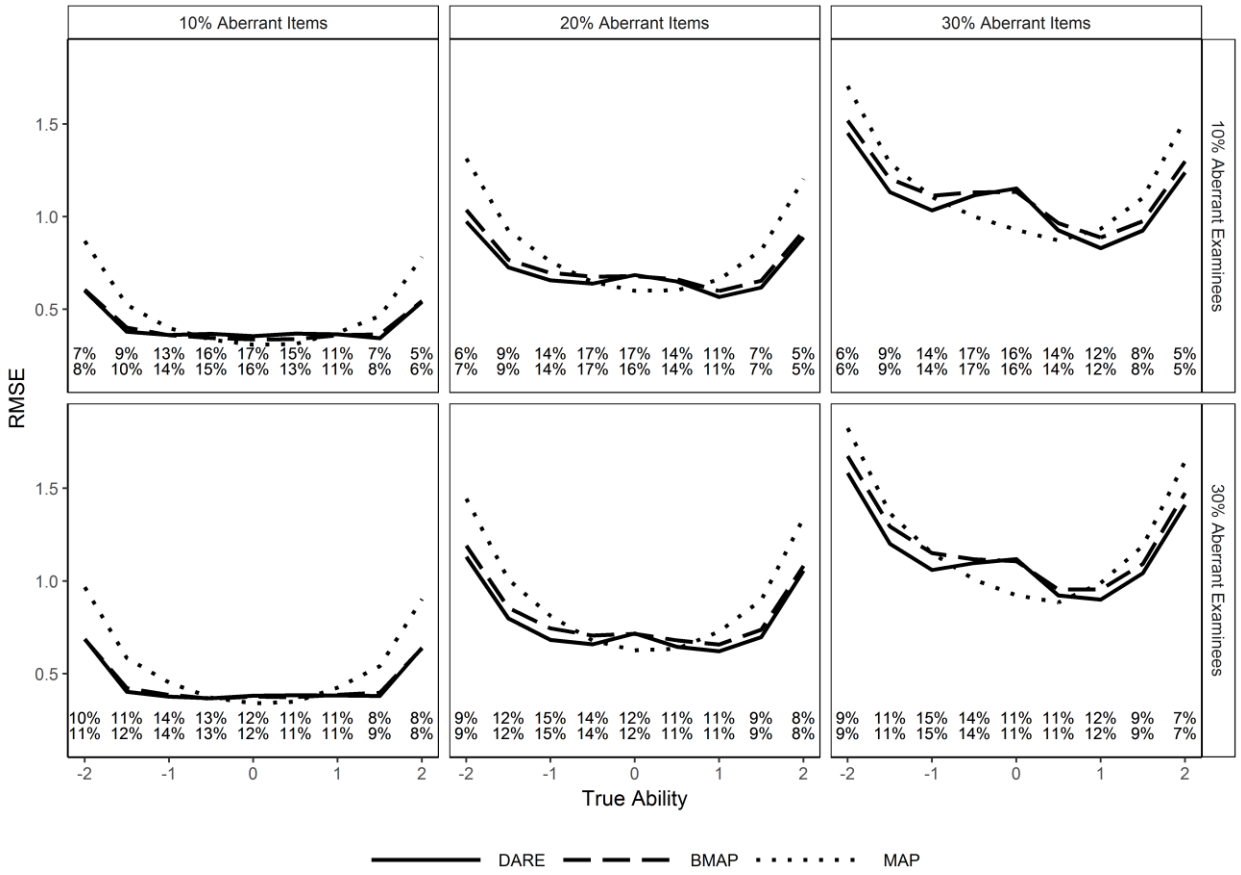

Figure 12. Ability estimation RMSE by θ (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Further, at all levels of MAP in the 60-item mixed behavior 10% aberrant items conditions, BMAP and DARE were associated with lower RMSE than MAP (see Figure 13). However, in the 20% and especially 30% aberrant items conditions, BMAP and DARE performed poorly compared to MAP among examinees with extreme MAP values. However, not many examinees had such extreme MAP estimates. In fact, almost all examinees in the 20% and 30% aberrant items conditions had MAP of about -0.5 to 0.5, and this was where it was most beneficial to use DARE instead of MAP and BMAP. The combination of Figures 12 and 13 showed that, for the mixed aberrant behavior situations, using DARE was the best choice to

obtain a low RMSE, especially among examinees with a θ distant from 0.0 and estimated MAP of near 0.0.

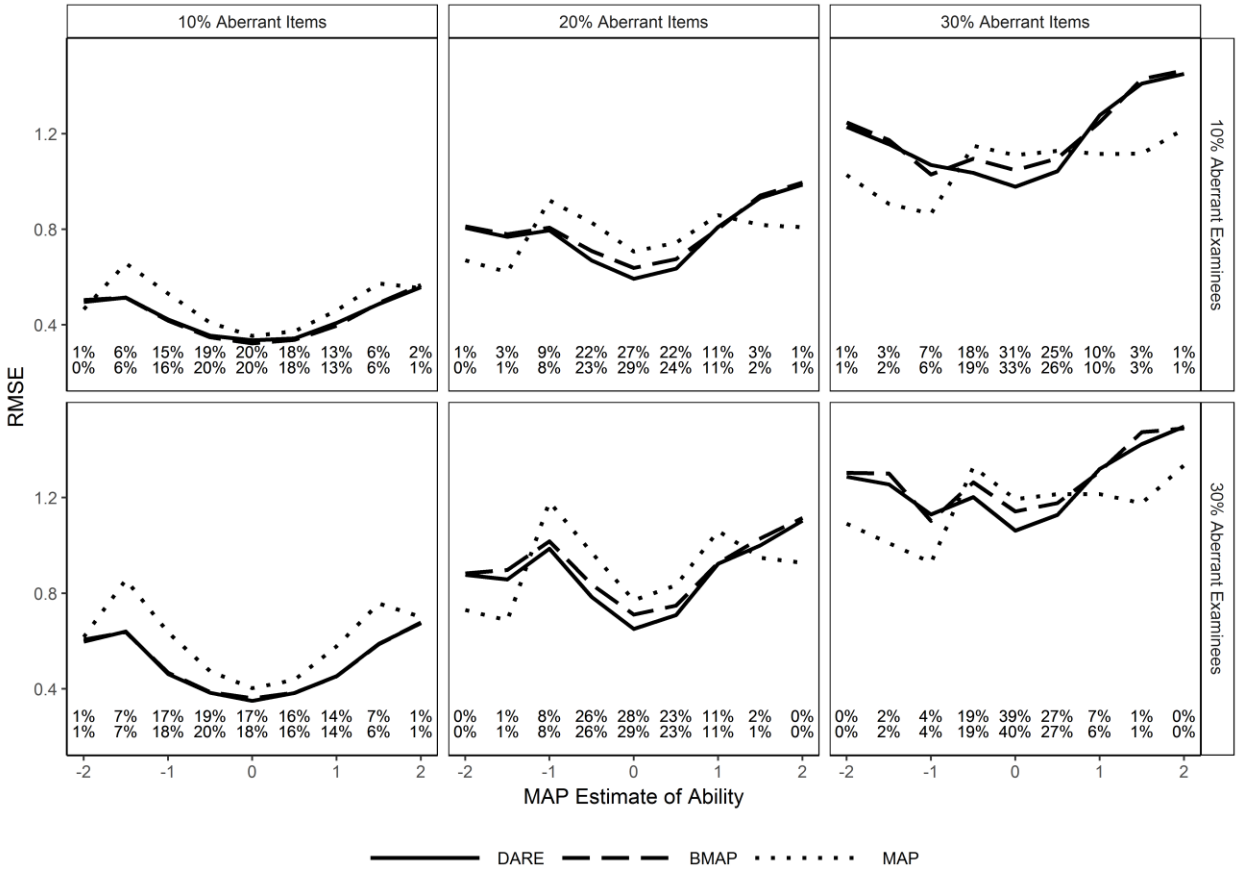

Figure 13. Ability estimation RMSE by MAP ability (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

The pattern of RMSE by θ with item difficulty parameters sampled from a $N(0,4)$ distribution rather than a $N(0,1)$ distribution showed increased effectiveness of using BMAP and DARE instead of MAP (see Figure 14). In addition, the advantage of using DARE instead of BMAP was extremely clear. Otherwise, the pattern observed in the comparable conditions remained the same.

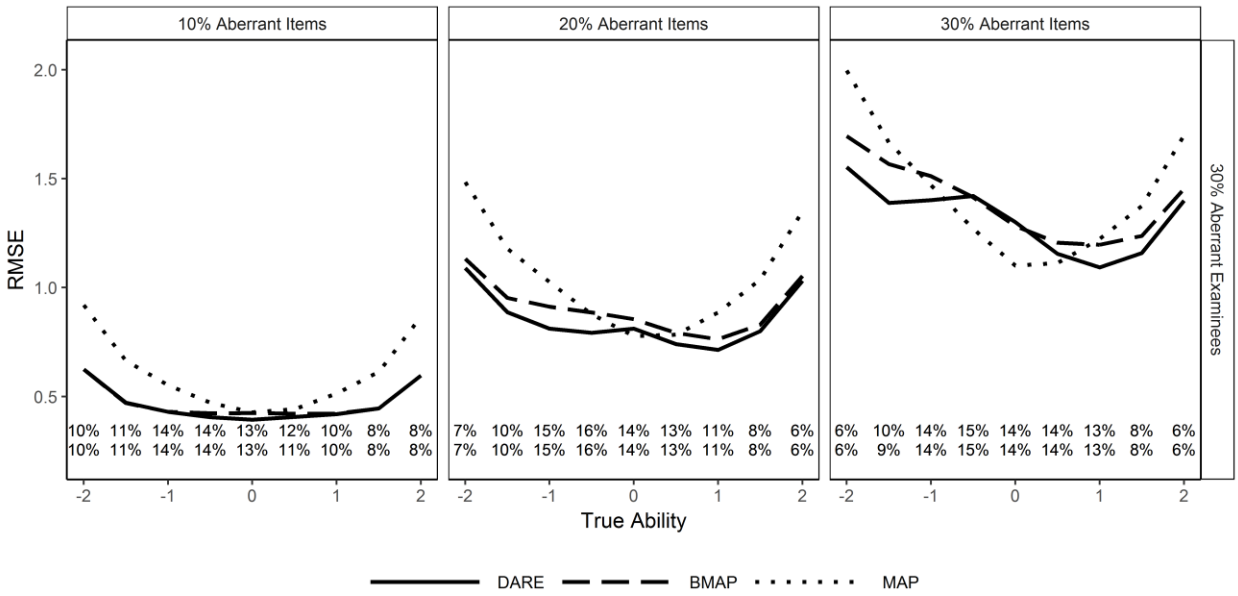

Figure 14. Ability estimation RMSE by θ with item difficulty variance of 4 (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Similarly, RMSE by MAP with item difficulty parameters sampled from a $N(0,4)$ distribution showed comparable patterns to the corresponding $N(0,1)$ distribution conditions but with an added effectiveness of using BMAP and DARE instead of MAP (see Figure 15). Again, the advantage of using DARE instead of BMAP was evident. Therefore, the combination of Figures 14 and 15 shows that the effectiveness of DARE over BMAP and MAP may increase when the item difficulty parameters have a flat distribution with relatively large numbers of very easy and very difficult items.

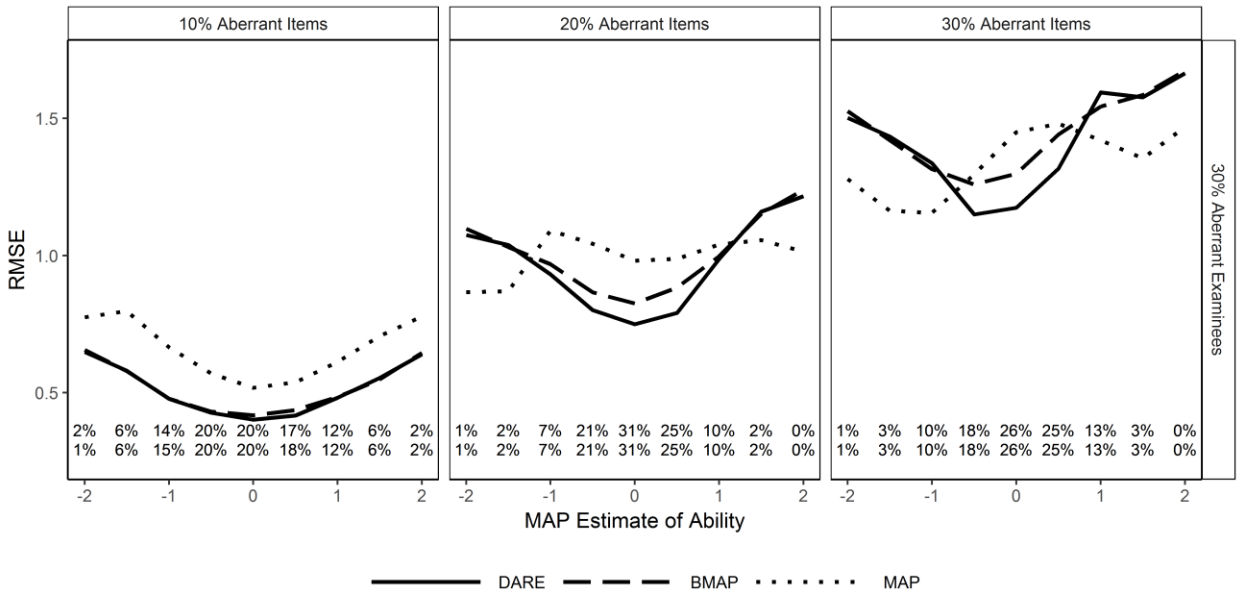

Figure 15. Ability estimation RMSE by MAP ability with item difficulty variance of 4 (60 items, mixed behavior). Examinees detected by l_z^* at the .01 level of α are included. Within each box, the top percentage values indicate the distribution of all detected examinees, while the bottom percentages are for the aberrant examinees.

Average DARE Weights

The weights w_i used in DARE to downweight misfitting items were examined for their effectiveness in downweighting items that were simulated as aberrant instead of non-aberrant. In the 20-item condition without aberrant behavior where l_z^* was used for detection, the average weight for the items was .92 (see Table 20). Similarly, in the conditions with aberrant behavior, the weights placed on the items among the non-aberrant examinees consistently stayed around .92 to .94. As expected, among the aberrant examinees, the aberrant items were always given a lower weight on average (i.e., .73 to .85) than non-aberrant items (i.e., .90 to .95). The difference in average weights between aberrant and non-aberrant items were slightly higher when the severity of aberrant behavior in the data set were low, such as the 10% aberrant examinees 10% aberrant items conditions. This shows that the aberrant items were easier to detect when there were not many aberrant items overall in the entire data set.

Table 20.
Average DARE Weights among Examinees Detected with l_z^* (20 Items)

AB	AE	AI	Aberrant Examinees		Non-Aberrant Examinees
			Aberrant Items	Non-Aberrant Items	
None	0%	0%			.92
SH	10%	10%	.74	.94	.92
		20%	.75	.93	.92
		30%	.80	.91	.92
	30%	10%	.80	.94	.93
		20%	.80	.93	.93
		30%	.85	.90	.94
SL	10%	10%	.74	.94	.92
		20%	.74	.94	.92
		30%	.78	.92	.92
	30%	10%	.79	.94	.92
		20%	.80	.93	.94
		30%	.83	.91	.94
Mix	10%	10%	.73	.94	.92
		20%	.73	.94	.92
		30%	.77	.92	.92
	30%	10%	.75	.95	.92
		20%	.76	.94	.93
		30%	.78	.93	.94

Note. DARE=downweighting of aberrant responses estimation; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

Increasing the test length from 20 to 40 items increased all DARE weights on average by about .01 to .02 (see Table 21). Otherwise, the relative pattern of the weights remained the same. Among the aberrant examinees, aberrant items continued to have lower weights (i.e., .75 to .85) compared to non-aberrant items (i.e., .92 to .96).

Table 21.
Average DARE Weights among Examinees Detected with l_z^* (40 Items)

AB	AE	AI	Aberrant Examinees		Non-Aberrant Examinees
			Aberrant Items	Non-Aberrant Items	
None	0%	0%			.93
SH	10%	10%	.76	.95	.94
		20%	.77	.94	.94
		30%	.81	.92	.94
	30%	10%	.80	.96	.95
		20%	.80	.94	.96
		30%	.85	.92	.96
SL	10%	10%	.76	.95	.94
		20%	.77	.95	.94
		30%	.79	.93	.94
	30%	10%	.81	.95	.94
		20%	.81	.95	.96
		30%	.83	.93	.96
Mix	10%	10%	.75	.95	.94
		20%	.75	.95	.94
		30%	.79	.93	.94
	30%	10%	.77	.96	.95
		20%	.77	.96	.96
		30%	.80	.94	.96

Note. DARE=downweighting of aberrant responses estimation; AE=aberrant examinees; AI=aberrant items; AB=aberrant behaviors; SL=spuriously low; SH=spuriously high

Similarly, increasing the test length from 40 to 60 items increased all DARE weights on both aberrant and non-aberrant items by about .01 in many conditions (see Table 22). Otherwise, the relative pattern of the weights remained the same. The exception was the 50% aberrant items conditions among the aberrant examinees where the aberrant items received higher weights than the non-aberrant items by .03. This showed that the high proportion of aberrant items caused the aberrant items to be classified as slightly more model fitting than the non-aberrant items. This may suggest that aberrance cannot be detected at all when the percentage of aberrant responses reach such severity. In addition, when the item difficulty parameters were sampled from a $N(0,4)$ distribution rather than $N(0,1)$, the difference between aberrant and non-aberrant items

increased by .04 to .07. Therefore, the aberrant items were easier to detect when the item difficulty values had a wider spread.

Table 22.
Average DARE Weights among Examinees Detected with l_z^* (60 Items)

σ^2	AB	AE	AI	Aberrant Examinees		Non-Aberrant Examinees
				Aberrant Items	Non-Aberrant Items	
1	None	0%	0%			.94
	SH	10%	10%	.77	.96	.95
			20%	.77	.95	.95
			30%	.81	.92	.95
		30%	10%	.80	.96	.96
			20%	.81	.95	.97
			30%	.84	.92	.97
	SL	10%	10%	.77	.96	.95
			20%	.77	.95	.95
			30%	.79	.94	.95
		30%	10%	.81	.96	.95
			20%	.81	.96	.96
			30%	.83	.94	.97
	Mix	10%	10%	.75	.96	.95
			20%	.76	.95	.95
			30%	.80	.93	.95
			50%	.91	.88	.95
		30%	10%	.78	.97	.96
			20%	.78	.96	.97
			30%	.81	.95	.97
			50%	.92	.89	.97
4	Mix	30%	10%	.71	.97	.95
			20%	.73	.97	.96
			30%	.78	.94	.96

Note. DARE=downweighting of aberrant responses estimation; σ^2=variance of the item difficulty parameter; AB=aberrant behaviors; AE=aberrant examinees; AI=aberrant items; SL=spuriously low; SH=spuriously high

Overall, these results indicate that the DARE procedure was identifying aberrant items as theoretically designed, where the aberrant items typically received weights that were about 10% to 20% lower than non-aberrant items. For the most part, the non-aberrant items received about the same degree of downweighting regardless of whether the examinee was truly aberrant.

CHAPTER 6: DISCUSSION

Most modern robust estimation methods downweight uninformative items regardless of the observed response. The rationale for such a practice is that when these responses are not aberrant, they provide little information about the latent trait level. However, if they are aberrant, they can cause large measurement error. Two drawbacks with such an approach is that 1) uninformative items still typically provide a substantial amount of information, and 2) not all uninformative responses are aberrant. These are especially true for non-aberrant or mildly aberrant examinees. This study presents the development of an aberrant response detection technique that is a complementary addition to the pre-existing robust estimation methods. The proposed robust ability estimation procedure, downweighting of aberrant responses estimation (DARE), downweights items based on the degree of model misfit as well as the amount of item information. This approach ensures that uninformative items with a high possibility of being aberrant are downweighted the most. This study also evaluates other popular robust and non-robust estimation methods, including maximum likelihood estimation (MLE), biweight (Mislevy & Bock, 1982), maximum a-posteriori (MAP), and biweight-MAP (BMAP; Maeda & Zhang, 2017b).

This study incorporates the person-fit analysis as a realistic initial step to identify the potentially aberrant examinees that might benefit from robust estimation methods. The person-fit detection results show that Type I error rates using both l_z^* and H^T are almost always deflated when aberrant behavior are present, and l_z^* has consistently higher power than H^T. The results indicate that the decrease in bias and RMSE by using DARE instead of MAP is about equal for l_z^* and H^T. However, given that l_z^* has consistently higher power and lower Type I error than H^T, l_z^* is superior to H^T because DARE can improve the ability estimate for more aberrant

examinees, while avoiding decreases in estimation accuracy among the non-aberrant examinees. The higher power of l_z^* than H^T is actually surprising as previous studies suggest the opposite (Karabastos, 2003; Tendeiro & Meijer, 2014). One possible reason is Karabastos's study was based on the Rasch model, while the current study used the 2-parameter logistic model (2PL). In addition, Tendeiro and Meijer (2014) reported higher Type I error rates for H^T (about .06 across all conditions at the .05 α level) than the current study, which may reflect differences in the methods used to identify the H^T cutoff value.

Based on l_z^*, DARE most effectively decreases the RMSE and bias among examinees detected at the .01 α cutoff. This is consistent with the threshold reported by Maeda and Zhang (2017b). Given such a low alpha level, the detection power of l_z^* is undesirably low, typically only detecting a third or less of the aberrant examinees. The effectiveness of DARE and all other robust estimation methods should improve if the person-fit analysis can detect more aberrant examinees. For example, in some testing situations, only certain forms of aberrant behavior may be suspected, such as copying (Wollack, 1997) or speededness (Shao & Cheng, 2015). Strategies specifically designed to detect those behaviors (e.g., the Omega statistic; Wollack, 1997) may be more powerful than generic person-fit statistics such as the l_z^*. In such cases, an appropriately modified version of DARE may have the potential to show larger effects than what was observed in this study.

Also, detection power may increase if robust methods are used to increase the accuracy of the ability estimate used in the parametric person-fit analysis. Some researchers have tried this with the biweight method and showed promising results (Meijer & Nering, 1997; Reise, 1995). Given that DARE was found to be more accurate than biweight under most examined conditions in this study, using DARE for initial person-fit analyses may be fruitful. Finally, the power of l_z^*

decreases as the percent of aberrant examinees increases, which may indicate inaccurate estimation of item parameters. Methods to improve item parameter estimation, such as removing misfitting examinees from the item parameter estimation process, may improve the power of person-fit analyses and ultimately improve the performance of DARE.

Overall, ability estimation accuracy based on bias and RMSE converges for most conditions. MAP is associated with the lowest bias and RMSE when no aberrant examinee is present. In general, DARE is the most accurate for tests with 60 items. At the 40-item level, DARE and BMAP have about equally the lowest RMSE, but DARE has slightly lower bias than BMAP. At the 20-item level, the effectiveness of all robust methods is questionable. Therefore, the usefulness of DARE depends on test length. Furthermore, DARE has the largest advantage over MAP in the mixed behavior conditions. Having a balance of both spuriously high and low scoring aberrant examinees may help item difficulty parameter estimation, which in turn helps DARE detect the aberrant items. Additionally, DARE works better for tests with larger spread of item difficulties (variance of 4 vs. 1). Practically, this means that a reasonable way to improve the person-fit analysis and robust estimation is to include a few extremely easy and hard items in the test.

When examined by the ability level, in the 60-item mixed aberrant behavior situations, DARE is more effective than MAP when applied to examinees with an extreme θ (i.e., less than -1.0 or more than 1.0) and estimated MAP of near 0.0 (i.e., -0.5 to 0.5). Also, in the spuriously high behavior condition, DARE is most effective when θ is very low (i.e., less than -1.0) and the MAP is between -1.0 and 0.0. Although not presented, the results for the spuriously high conditions are similar to that for the spuriously low conditions. These findings show that DARE performs well when 1) a high-achieving person obtains spuriously incorrect responses, which

results in no less than about an average ability estimate, or 2) a low-achieving person obtains spuriously correct responses, which results in no more than about an average ability estimate. Cases like this with mild aberrant behavior seem to be commonly occurring (Rupp, 2013). On the other hand, DARE seems to have a difficult time dealing with the aberrant response patterns from extremely high or low ability estimates (e.g., MAP of less than -2.0 or over 2.0), regardless of the true ability. These findings have important implications in test uses, such as proficiency classification. For example, in a context where cheating is suspected, DARE will not be effective in improving classification accuracy if the cut score is set high, such as the 90th percentile. However, at the 50th percentile cutoff, DARE will be extremely useful in preventing underachieving cheaters from passing the test.

Finally, to shed some light on the nature of the aberrant behavior, one may examine the weights that DARE applies to each item. In this study, aberrant items have received weights that were about 10% to 20% lower than non-aberrant items. This is true even in the 20-item conditions where the performance of DARE is not always superior to other methods. In other words, as theoretically designed, DARE is capable of distinguishing between aberrant and non-aberrant responses.

Limitations

This simulation study has been designed to be as realistic as possible by using estimated item parameters and conducting person-fit analyses to detect examinees for robust ability estimation. However, item analyses were not performed. In practice, low quality items should be identified and may be removed. Also, although the items in the study were generated based on the 2PL model, the 3PL model may fit better for some items after the aberrance was added, such

as those with many spuriously incorrect responses. Taking this extra step prior to person-fit analyses may alter the effectiveness of DARE.

Further, like any simulation study, the tests simulated in the current study are relatively simple in that the non-aberrant responses fit a unidimensional 2PL model perfectly. In reality, tests often measure more than one dimension, use polytomous items and testlets, and have missing responses. Extension of the robust estimation methods to these more realistic conditions are necessary. Other conditions not explored in this study include aberrant behaviors that cause both spuriously correct and incorrect responses (i.e., spuriously mixed responses; Rupp, 2013), behaviors that only affect items of a certain difficulty, cases where the item parameters have been pre-calibrated, or cases where additional useful information about the examinees is available (e.g., scores on a similar test).

Finally, a potential problem with DARE and other robust estimation methods is that downweighting some items may result in a change in the test content. For example, a unidimensional mathematics test can contain equal numbers of algebra and geometry items, and DARE may end up downweighting algebra items more so than geometry. In this case, the resulting mathematics ability estimate may be more representative of geometry than the equal combination of geometry and algebra. The chance that this will occur may increase if one section is more difficult than the other. Possible solutions include 1) constraining DARE to downweight each sub content area equally, 2) upweighting some well-fitting items in the sub content area that received heavy downweighting, or 3) reporting the ability estimate within each sub content area.

Conclusion

Test responses can be contaminated by aberrant behaviors. Ignoring such behaviors can result in inaccurate test scores and score interpretations for both aberrant and non-aberrant test takers. Effective techniques that can identify aberrant examinees and correct their test scores are immensely valuable. So far, much of the literature on aberrant behaviors has focused on how to detect the aberrant response patterns. The current study provides a better solution to correct contaminated test scores.

REFERENCES

Armstrong, R. D., & Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*, 408-428.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society. Series B, 50*, 225-265.

Cui, Y., & Leighton, J. P. (2009). The Hierarchy Consistency Index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46,* 429-449.

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159-177.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191.

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Emons, W. H. M. (2003). Investigating the local fit of item-score vectors. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.). *New developments in psychometrics* (pp. 289-296). Tokyo: Springer.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*(1), 101-119.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159-180.

Guttman, L. (1944). A basis for scaling qualitative data. American *Sociological Review*, *9*, 139-150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A Claussen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.

Lee, P., Stark, S., & Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise Preference tests: An application of lz based on the Zinnes-Griggs ideal point IRT model. *Applied Psychological Measurement, 38*(5), 391-403.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269-290.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579-598.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*, 477-482.

Maeda, H., & Zhang, B. (2017a). An algorithm to improve test answer copying detection using the omega statistic. *International Journal of Testing*, *17*(1), 55-73.

Maeda, H., & Zhang, B. (2017b). *Bayesian extension of biweight and Huber weight for robust ability estimation.* Manuscript submitted for publication.

Magis, D., Raiche, G., & Beland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57-81.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, *21*, 321-336.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimators of latent ability. *Educational and Psychological Measurement*, *42*, 725-737.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75-106.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*, 121-129.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, *22*, 53-69.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213-229.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543-570.

Reise, S. P., & Due, A. M. (1991). Test characteristics and their influence on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217–226.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17* (5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3-38.

Schuster, C., & Yuan, K-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*(6), 720-735

Shao, C., Li, J. & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*(4), 1118 –1141.

Sinharay, S. (2016a). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of Mathematical and Statistical Psychology*, *69*, 175-193.

Sinharay, S. (2016b). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement, 53,* 63-85.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, *45*, 433-444.

Snijders, T. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342.

Strandmark, N. L. & Linn, R. L. (1987). A generalized logistic item response model parametrizing test score inappropriateness. *Applied Psychological Measurement*, *11*, 355-370.

Tendeiro, J. N. (2015). *PerFit: Person Fit*. R package version 1.4. https://CRAN.R-project.org/package=PerFit

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83–108). New York: Academic Press.

Wainer, H. & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*(3), 373-391.

Waller, M. I. (1974). *Removing the Effects of Random Guessing from Latent Trait Ability Estimates.* Educational Testing Service, Princeton N.J. ETS-RB-74-32.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, *21*, 307-320.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466-479.

APPENDIX: R Code to Calculate DARE

The R function `DARE` is presented. This function uses item response data and item parameters and outputs the DARE ability estimates and the weights. The function `BMAP` is presented for calculations of the BMAP estimate because this is required when using `DARE`. Using the default settings are recommended. Missing data are treated as missing completely at random and their effects on the likelihood function are disregarded. Finally, an example is presented.

Arguments

`DAT`	Person by item data.frame or matrix.
`ITEMPAR`	Item by parameter 2-parameter logistic model parameter matrix. Item discrimination is in the first column, and item difficulty is in the second column.
`crit.lzw`	Conversion criterion for global lzw. Defaults to -1.645
`TuCo`	Tuning coefficient. Defaults to 5.
`crit.newton`	Newton-Raphson convergence criterion. Defaults to .0001.
`prior.mean`	Bayesian prior normal distribution mean. Defaults to 0.
`prior.std`	Bayesian prior normal distribution standard deviation. Defaults to 1.
`max.iter1`	Maximum number of DARE iterations. Defaults to 99999.
`max.iter2`	Maximum number of Newton-Raphson iterations. Defaults to 25.
`Increment`	Degree of downweighting in every DARE iteration. Defaults to .1.
`limit.upper`	Upper limit of ability estimates. Defaults to 4.
`limit.lower`	Lower limit of ability estimates. Defaults to -4.

Value

`Theta`	Vector of DARE ability estimate for each person.
`Weight`	Person by item DARE weight matrix.

Code

```
#################################################################
### Create DARE Function
DARE <- function(
  DAT,
  ITEMPAR,
  crit.lzw = -1.645,
  TuCo = 5,
  crit.newton = .0001,
  prior.mean = 0,
  prior.std = 1,
  max.iter1 = 99999,
  max.iter2 = 25,
  Increment = .1,
  limit.upper = 4,
  limit.lower = -4){

  ### Create a Function to Calculate lz for Subtests
  LZcalc <- function(x){
    (sum(x[,1])-sum(x[,2]))/sum(x[,3])^.5
    }

  ### Get Number of Columns and Rows
  k <- ncol(DAT)
  N <- nrow(DAT)

  ### Initialize Output
  Out <- rep(NA,N)
  Weight <- matrix(NA,nrow = N, ncol = k)

  ### Initial BMAP Estimate
  EST <- BMAP(
    DAT = DAT,
    ITEMPAR = ITEMPAR,
    EST = rep(0,nrow(DAT)),
    TuCo = 4,
    prior.mean = prior.mean,
    prior.std = prior.std,
    crit.newton = crit.newton,
    max.iter = max.iter2,
    limit.upper = limit.upper,
    limit.lower = limit.lower)

  ### Make 3 Subtests of About the Same Length
  len <- round(k/3)
```

```
SubtestNum <- rep(2,k)
SubtestNum[1:len] <- 1
SubtestNum[(k-len+1):k] <- 3

### Begin Person Loop
for(i in 1:N){
  if(is.na(EST[i])) next
  u0 <- unlist(DAT[i,])
  theta0 <- EST[i]
  theta1 <- EST[i]
  w <- rep(1,k)

  ### Begin DARE Loop to Determine w
  for(j in 1:max.iter1){
    Z <- ITEMPAR[,1]*(theta0-ITEMPAR[,2])
    ppp1 <- 1/(1+exp(-Z))
    ppp0 <- 1-ppp1
    ppp <- ppp1
    ppp[u0 == 0 & !is.na(u0)] <- 1-ppp[u0 == 0 & !is.na(u0)]

    #Set Missing Data Probabilities to 1.
    #This Removes Their Effects on the Likelihood
    ppp[is.na(u0)] <- 1
    ppp0[is.na(u0)] <- 1
    ppp1[is.na(u0)] <- 1
    ORDER <- rev(order(ppp1))

    ### Find 1 Misfitting Subtest
    l0 <- log(ppp)*w
    Exp <- log(ppp1)*ppp1*w+log(ppp0)*ppp0*w
    Var <- ppp1*ppp0*w^2*(log(ppp1/ppp0)^2)
    lzwSubtests <- c(by(cbind(l0,Exp,Var)[ORDER,],
      SubtestNum,LZcalc))
    SelectSubtest <- which.min(lzwSubtests)

    ### Find 1 Misfitting Item
    ItemSet <- ORDER[c(1:k)[SubtestNum == SelectSubtest]]
    ItemSet <- ItemSet[ppp[ItemSet] < .5 & w[ItemSet]>.01]
    SelectItem <- ItemSet[which.min(ITEMPAR[ItemSet,1])]

    ### Downweight Item
    w[SelectItem] <- round(w[SelectItem]-Increment,3)

    ### Begin BMAP Newton-Raphson loop
    #Remove Effects of Missing Data on the Likelihood
    for(l in 1:max.iter2){
      Z <- ITEMPAR[,1]*(theta0-ITEMPAR[,2])
```

```
    ppp1 <- 1/(1+exp(-Z))
    r <- Z
    g <- w*(1-(r/TuCo)^2)^2
    g[abs(r) > TuCo]=0
    num <- sum(c(g*ITEMPAR[,1]*(u0-ppp1))[!is.na(u0)]) -
      (theta0-prior.mean)/prior.std^2
    den <- -1*sum(c(g*ITEMPAR[,1]^2*ppp1*(1-
      ppp1))[!is.na(u0)])- 1/prior.std^2
    theta1 <- theta0-num/den

    ### Convergence Criteria
    if(abs(theta1-theta0) < crit.newton|is.na(theta1)) break
    if(theta1 > limit.upper) {
      theta1 <- limit.upper
      break
    }
    if(theta1 < limit.lower) {
      theta1 <- limit.lower
      break
    }
  theta0 <- theta1
  } # End BMAP Newton-Raphson loop

  ### Find Global lzw
  Z <- ITEMPAR[,1]*(theta0-ITEMPAR[,2])
  ppp1 <- 1/(1+exp(-Z))
  ppp0 <- 1-ppp1
  ppp <- ppp1
  ppp[u0 == 0 & !is.na(u0)] <- 1-ppp[u0 == 0 & !is.na(u0)]
  #Set Missing Data Probabilities to 1.
  #This Removes Their Effects on the Likelihood
  ppp[is.na(u0)] <- 1
  ppp0[is.na(u0)] <- 1
  ppp1[is.na(u0)] <- 1
  l0 <- sum(log(ppp)*w)
  Exp <- sum(log(ppp1)*ppp1*w+log(ppp0)*ppp0*w)
  Var <- sum(ppp1*ppp0*w^2*(log(ppp1/ppp0)^2))
  lzw <- (l0-Exp)/Var^.5

  ### Assess Global Person-Fit
  if(lzw > crit.lzw | is.na(lzw)) break
} # End DARE loop to determine w

### Save Output
Out[i] <- theta0
Weight[i,] <- w
} # End Person Loop
```

```
  return(list(Theta = Out, Weight = Weight))
}

#################################################################
### Create BMAP Function
BMAP <- function(
  DAT,
  ITEMPAR,
  EST = rep(0,nrow(DAT)),
  TuCo = 4,
  prior.mean = 0,
  prior.std = 1,
  crit.newton = .0001,
  max.iter = 25,
  limit.upper = 4,
  limit.lower = -4){

  ### Get Number of Columns and Rows
  k <- ncol(DAT)
  N <- nrow(DAT)

  ### Initialize Output
  Out <- rep(NA,N)

  for(i in 1:N){
    u0 <- unlist(DAT[i,])
    theta0 <- EST[i]

    ### Begin BMAP Newton-Raphson Loop
    #Remove Effects of Missing Data on the Likelihood
    for(l in 1:max.iter){
      Z <- ITEMPAR[,1]*(theta0-ITEMPAR[,2])
      ppp1 <- 1/(1+exp(-Z))
      r <- Z
      g <- (1-(r/TuCo)^2)^2
      g[abs(r) > TuCo] <- 0
      num <- sum(c(g*ITEMPAR[,1]*(u0-ppp1))[!is.na(u0)]) -
        (theta0-prior.mean)/prior.std^2
      den <- -1*sum(c(g*ITEMPAR[,1]^2*ppp1*
        (1ppp1))[!is.na(u0)])- 1/prior.std^2
      theta1 <- theta0-num/den

      ### Convergence Criteria
      if(abs(theta1-theta0)<crit.newton|is.na(theta1))  break
      if(theta1 > limit.upper) {
        theta1 <- limit.upper
```

```
        break
      }
      if(theta1 < limit.lower) {
        theta1 <- limit.lower
        break
      }

      ### Save New Theta Value
      theta0 <- theta1
    } # End BMAP Newton-Raphson Loop

    Out[i] <- theta1
  } #End Person Loop
  return(Out)
}

#################################################################
### EXAMPLE

#Generate Parameters
set.seed(71690)
Npeople <- 10
Nitems <- 10
TrueAbility <- rnorm(Npeople)
itempar <- matrix(c(rlnorm(Nitems,.4,.5),rnorm(Nitems)),ncol=2)
colnames(itempar) <- c("a","b")

#Create Data
dat <- matrix(nrow=Npeople,ncol=Nitems)
for(i in 1:Npeople){
  for(j in 1:Nitems){
    Z <- itempar[j,1]*(TrueAbility[i]-itempar[j,2])
    p <- 1/(1+exp(-Z))
    dat[i,j] <- sample(c(1,0),size=1,prob=c(p,1-p))
    }
  }

#Run DARE
out <- DARE(DAT=dat,ITEMPAR=itempar)

#DARE Ability Estimates
out$Theta

#DARE Weights
out$Weight
```

CURRICULUM VITAE

Education

Ph.D., Educational Psychology, Educational Statistics & Measurement Concentration
University of Wisconsin-Milwaukee, Milwaukee, WI (August 2017)
Dissertation: Robust latent ability estimation based on item response information and model fit

M.S., Exercise and Sport Science, Physical Activity Promotion Concentration
East Carolina University, Greenville, NC (August 2013)
Thesis: Introducing portable pedal machines inside a university library to reduce sedentary behavior

B.S., Kinesiology
Washington State University, Pullman, WA (May 2011)

Publications

1. Maeda, H., & Zhang, B. (under review). Bayesian extension of biweight and Huber weight robust latent ability estimation.

2. Rowley, T. W., Lenz, E. K., Swartz, A. M., Miller, N. E., Maeda, H., & Strath, S. J. (under review). Efficacy of an individually-tailored, internet-mediated physical activity intervention in older adults: A randomized controlled trial. *Journal of Applied Gerontology*.

3. Kim, S-Y., Fouad, N., Maeda, H., Xie, H., & Nazan, N. (2017). Mid-life work and psychological well-being: A test of the psychology of working theory. *Journal of Career Assessment*. DOI: DOI:10.1177/1069072717714538

4. Kavanaugh, M. S., Cho, C., Maeda, H., & Swope, C. (2017). "I am no longer alone": Evaluation of the first North American camp for youth in families with Huntington's Disease. *Children and Youth Services Review*, *79*, 325-332.

5. Choi, A. Y., Israel, T., & Maeda, H. (2017). Development and evaluation of the Internalized Racism in Asian Americans Scale (IRAAS). *Journal of Counseling Psychology, 64*(1), 52-64.

6. Maeda, H., & Zhang, B. (2017). An algorithm to improve test answer copying detection using the omega statistic. *International Journal of Testing*, *17*(1), 55-73.

7. Kwak, J., Brondino, M. J., O'Connell Valuch, K., & Maeda, H. (2016). Evaluation of the Music and Memory program among nursing home residents with dementia: Final Report to the Wisconsin Department of Health Services. Retrieved from https://www.dhs.wisconsin.gov/publications/p01594.pdf

8. Lerma, N., Swartz, A. M., Rowley, T., Maeda, H., & Strath, S. (2016). Increasing the energy expenditure of seated activities in older adults with a portable elliptical device. *Journal of aging and physical activity*, 25, 99 -104.

9. Quartiroli, A., & Maeda, H. (2016). The effects of a lifetime physical activity and fitness course on college students' health behaviors. *International Journal of Exercise Science, 9*(2), 136-148.

10. Cole, Z., & Maeda. H. (2015). Effects of listening to preferential music on sex differences in endurance running performance. *Perceptual & Motor Skills, 121*(2), 390-398.

11. Maeda. H., Quartiroli, A., Vos, P., Carr, L. J., & Mahar, M. T. (2014). Feasibility of retrofitting a university library with active workstations to reduce sedentary behavior. *American Journal of Preventive Medicine, 46*(5), 525-528.

12. Maeda, H. (2014). Response option configuration of online administered Likert scales. *International Journal of Social Research Methodology, 18*(1), 15-26. DOI:10.1080/13645579.2014.885159

13. Swartz, A., Rote, A., Welch, W., Maeda, H., Hart, T., Cho, Y., & Strath, S. (2014). Prompts to disrupt sitting time and increase physical activity at work. *Preventing Chronic Disease, 11*. DOI: http://dx.doi.org/10.5888/pcd11.130318

14. Quartiroli, A., & Maeda, H. (2014). Self-determined engagement in physical activity and sedentary behaviors of US college students. *International Journal of Exercise Science, 7*(1), 87-97. Full text: http://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1525&context=ijes

15. Carr, L. J., Maeda, H., Luther, B., Rider, P., Tucker, S., & Leonhard, C. (2014). Acceptability and effects of a seated active workstation during sedentary work: A proof of concept study. *International Journal of Workplace Health Management, 7*(1), 2-15.

Conference Presentations

1. Wang, S., Maeda, H., & Zhang, B. (2017). *Understanding PISA problem solving assessment: Cross-country and cross-race comparisons.* National Council on Measurement in Education Annual Meeting, San Antonio, TX. Poster presentation

2. Bordon, J. J., & Maeda, H. (2016). *Transitions of ethnic label use and its effect on mental health for Asian adolescents in America.* 124th Annual Meeting of the American Psychological Association, Denver, CO. Poster presentation

3. Choi, A. Y., Israel, T., & Maeda, H. (2016). *Construct validation of the Internalized Racism in Asian Americans Scale (IRAAS).* 124th Annual Meeting of the American Psychological Association, Denver, CO. Poster presentation

4. Maeda, H., Wang, S., & Azen, R. (2016). *Comparison of two methods for determining the number of factors in exploratory factor analysis.* 124th Annual Meeting of the American Psychological Association, Denver, CO. Poster presentation

5. Maeda, H., & Zhang, B. (2016). *An iterative technique to improve test cheating detection using the omega statistic.* National Council on Measurement in Education Annual Meeting, Washington DC. Poster presentation

6. Xie, H., Maeda, H., & Kim, S-Y. (2016). *Smoking behavior, sexual orientation, and depression among U.S. adults: 2005–2012 NHANES.* 124th Annual Meeting of the American Psychological Association, Denver, CO. Poster presentation

7. Choi, A. Y., Maeda, H., & Bordon, J. J. (2016). *Measurement Invariance of the Internalized Racism in Asian Americans Scale (IRAAS).* Annual Meeting of the Asian American Psychological Association, Denver, CO. Poster presentation

8. Maeda, H., Cho, Y., & Cho, C. (2016). *Psychometric Analysis of the 6-Dimensional Community Characteristics Scale.* College of Health Sciences Symposium, University of Wisconsin-Milwaukee, Milwaukee, WI. Poster presentation

9. Xie, H., Maeda, H., Kim, S-Y. (2015). *Psychological well-being in lesbian, gay men, and bisexuals: The effect of perceived discrimination in the United States.* American Public Health Association Annual Meeting & Convention, Chicago, IL. Poster presentation

10. Kim, S-Y., Maeda, H., Xie, H. (2015). *Paternal involvement, men's work family spillover and psychological well-being.* American Psychological Association Annual Convention, Toronto, Canada. Poster presentation

11. Kim, S-Y., Fouad, N., Maeda, H., Xie, H. (2015). *Midlife people's work and psychological well-being: The psychology of working perspective.* American Psychological Association Annual Convention, Toronto, Canada. Poster presentation

12. Strath, S. J., Cho, Y. I., Welch, W. A., Maeda, H., Rowley, T. W., Miller, N. E., & Swartz, A. M. (2015). *Simulation of accelerometer data reduction choices on sample size and select physical activity and sedentary outcomes in older adults.* International Conference on Ambulatory Monitoring of Physical Activity and Movement, Limerick, Ireland. Poster presentation

13. Lerma, N. L., Swartz, A. M., Rowley, T. W., Maeda, H., & Strath, S. J. (2015). *Increasing energy cost of sedentary behaviors in older adults using a portable elliptical device: Pilot examination.* American College of Sports Medicine, National Meeting, San Diego, CA. Poster presentation

14. Maeda, H., Zhang, B. (2015). *An iterative technique to detect test cheating using the nominal response model.* School of Education Research Gala, University of Wisconsin-Milwaukee, Milwaukee, WI. Poster presentation

15. Lerma, N. L., Swartz, A. M., Rowley, T. W., Maeda, H., & Strath, S. J. (2015). *Increasing energy cost of sedentary behaviors in older adults using a portable elliptical device: Pilot examination.* College of Health Sciences Symposium, University of Wisconsin-Milwaukee, Milwaukee, WI. Poster presentation

16. Maeda, H. & Lerma, N. L. (2014). *Sedentary behavior: Promoting regular exercise is not enough.* Wisconsin Public Health Association Annual Conference, Milwaukee, WI. Poster presentation

17. Reesor, L. Maeda, H., Raedeke, T. D., Gross McMillan, A., & DuBose, K. D. (2014). *Barriers & physical activity participation among normal weight and overweight/obese children*. American College of Sports Medicine Annual Meeting, Orlando, FL. Poster presentation

18. Maeda, H., Quartiroli, A., Vos, P. W., Carr, L. J., Mahar, M. T. (2014). *Feasibility of Retrofitting a Library to Reduce Sedentary Behavior*. American College of Sports Medicine Annual Meeting, Orlando, FL. Poster presentation

19. Mahar, M. T., Maeda, H., Sung, H., Mahar, T. F. (2014). *Accuracy of the Nike Fuelband+ and Fitbit One Activity Monitors*. American College of Sports Medicine Annual Meeting, Orlando, FL. Poster presentation

20. Maeda, H., Quartiroli, A., Vos, P. W., Carr, L. J., Mahar, M. T. (2014). *Feasibility of Retrofitting a Library to Reduce Sedentary Behavior*. College of Health Sciences Spring Research Symposium, University of Wisconsin-Milwaukee, Milwaukee, WI. Poster presentation

21. Lerma, N., Maeda, H., Garcia, B., Trepasso, T., Bork, E., & Strath, S. J. (2014). *Multiple health-based point-of-decision prompts and an elevator survey to alter stair and elevator use.* College of Health Sciences Spring Research Symposium, University of Wisconsin-Milwaukee, Milwaukee, WI. Poster presentation

22. McKinnis, D., Webb, R., Maeda, H., DuBose, K. D., McMillan, A. G. (2014). *Body mass and motor skill performance in children 7-9 years*. Presented at Research & Creative Achievement Week, East Carolina University, Greenville, NC. Poster presentation (received Best Graduate Student Poster Award)

23. Reesor, L. Maeda, H., Raedeke, T. D., Gross McMillan, A., & DuBose, K. D. (2014). *Barriers & physical activity participation among normal weight and overweight/obese children*. Southeast Chapter of the American College of Sports Medicine Annual Meeting, Greenville, SC. Poster presentation

24. McKinnis, D., Webb, R., DuBose, K. D., Maeda, H., McMillan, A. G. (2013). *Body mass and motor skill performance in children 7-9 years.* Presented at 2013 North Carolina Physical Therapy Association Annual Fall Conference and Chapter Meeting, Asheville, NC. Poster presentation

25. Maeda, H. & Mahar, M. (2013). *What is the best response option configuration in online-administered Likert scales?* Presented at Research & Creative Achievement Week, East Carolina University, Greenville, NC. Oral presentation

26. Maeda, H., & Quartiroli, A. (2013). *Reducing sedentary behavior in a university library*. Presented at 2013 Annual Meeting Southeast Chapter of the American College of Sports Medicine, Greenville, SC. Poster presentation

27. Maeda, H., Cole, Z., & Brooks, J. (2011). *Music and extraversion on free throw pace*. Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Oral presentation

28. McCord, T. J., Bruya, L. D., Isom, J., McCullough, L., & Maeda, H. (2011). *Writing revision in kinesiology*. Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Oral presentation

29. Burns, A., Comfort, G., Maeda, H., Cole, Z., McCord, T., Clifton, T., Dotson, C., Kendall, G., Silvers, M., & Bruya, L. (2011). *Kinesiology student performance and instruction in the classroom.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Oral presentation

30. Van Mullen, H., Lawrence, D., Maeda, H., & Stoll, S. (2011). *Coaching: Is it for you?* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Panel discussion

31. Burns, A., Maeda, H., Dotson, C., Cole, Z., Catalano, L., Blehm, A., Sorensen, C., McGowan, J., Silvers, W. M., & Bruya, L. D. (2011). *A symposium on student performance outside the classroom: Student conferencing used for idea exchange*. Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Symposium presentation

32. Cole, Z., Maeda, H., Dotson, C., Arndt, S., Silvers, W. M., & Bruya, L. D. (2011). *A symposium on student performance outside the classroom: Professional review teams and student journal in undergraduate school*. Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Symposium presentation

33. Maeda, H. (2011). *Coed flag football: The Diamond offense.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

34. Cole, Z., Maeda, H., & Knutson, D. (2011). *Music selection and running performance.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

35. Maeda, H. (2011). *Group cohesion in flag football.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

36. Maeda, H., Cole, Z., & Brooks, J. (2011). *Music, distraction, and free throw accuracy.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

37. Maeda, H., Cole, Z., & Knutson, D. (2011). *Frequency of music listening.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

38. Maeda, H., Cole, Z., Boice, J., Swihart, R., & Mendes, K. (2011). *Music genre preference: Running vs studying.* Presented at the 56th Annual Western Society for Kinesiology & Wellness, Reno, NV. Poster presentation

39. Maeda, H., Cole, Z., Swihart, R., Boice, J., & Mendes, K. (2011). *Slow relaxing music and runners*. Presented at the 7th annual Northwest Student Professional Network Conference, Washington State University, Pullman, WA. Poster presentation

40. Maeda, H., Cole, Z., Swihart, R., Boice, J., & Mendes, K. (2011). *Slow relaxing music and runners*. Presented at Undergraduate Research Symposium, Washington State University, Pullman, WA. Poster presentation

41. Maeda, H. (2009). *Aging of skeletal muscles.* A review of Frontera, W.R., Hughes, V. A., Fielding, R. A., Fiatarone, M. A., Evans, W. J., & Roubenoff, R. (2000). Aging of skeletal muscle: A 12-yr longitudinal study. Presented at Geoff Wood Foundation Conference, Washington State University, Pullman, WA. Poster presentation

Graduate Research Assistantships

Data Manager, Center for Aging and Translational Research
University of Wisconsin-Milwaukee, Milwaukee, WI (March 2015-Present)

Research Assistant for Drs. Scott Strath & Ann Swartz, Department of Kinesiology
University of Wisconsin-Milwaukee, Milwaukee, WI (August 2013-August 2015)

Research Assistant for Dr. Matthew Mahar, Department of Kinesiology
East Carolina University, Greenville, NC (May 2013 to June 2013)

Research Assistant for Dr. Katrina DuBose, Department of Kinesiology
East Carolina University, Greenville, NC (January 2013 to May 2013)

Research Assistant for Dr. Ale Quartiroli, Department of Kinesiology
East Carolina University, Greenville, NC (August 2012 to December 2012)

Research Assistant for Dr. Lucas Carr, Department of Kinesiology
East Carolina University, Greenville, NC (January 2012 to May 2012)

Teaching Experience

Grader, ED PSY 325 Practice of Classroom Assessment
University of Wisconsin-Milwaukee, Milwaukee, WI (August 2014-December 2014)

Instructor, EXSS 1000 Lifetime Physical Activity and Fitness Laboratory
East Carolina University, Greenville, NC (Spring 2012 to Spring 2013)

Teacher's Assistant, EXSS 4501 Independent Study in Exercise and Sport Science
East Carolina University, Greenville, NC (Fall 2012)

Teacher's Assistant, MVTST 313 Behavioral Aspects of Human Movement
Washington State University, Pullman, WA (Spring, 2011)

Teacher's Assistant, Northwest Student Professional Network Publication of Works Review Team
Washington State University, Pullman, WA (Fall, 2010; Spring 2011)

Teacher's Assistant, MVTST 199 Human Motor Development
Washington State University, Pullman, WA (Fall, 2009)

Additional Research Experience

Lab assistant, Physical Activity and Health Research Lab
University of Wisconsin-Milwaukee, Milwaukee, WI (August 2014-August 2015)

Statistical Consultant, American Kinesiology Association (2013)

Biostatistics Internship with Dr. Paul Vos, Department Chair, Department of Biostatistics
East Carolina University, Greenville, NC (May 2012 to June 2012)