PSYCHOMETRIKA—VOL. 86, NO. 2, 464–488
JUNE 2021
https://doi.org/10.1007/s11336-021-09756-3

ASYMPTOTICALLY CORRECTED PERSON FIT STATISTICS FOR MULTIDIMENSIONAL CONSTRUCTS WITH SIMPLE STRUCTURE AND MIXED ITEM TYPES

MAXWELL HONG

UNIVERSITY OF NOTRE DAME

LIZHEN LIN

UNIVERSITY OF NOTRE DAME

YING CHENG

UNIVERSITY OF NOTRE DAME

Person fit statistics are frequently used to detect aberrant behavior when assuming an item response model generated the data. A common statistic, l_z, has been shown in previous studies to perform well under a myriad of conditions. However, it is well-known that l_z does not follow a standard normal distribution when using an estimated latent trait. As a result, corrections of l_z, called l_z^*, have been proposed in the literature for specific item response models. We propose a more general correction that is applicable to many types of data, namely survey or tests with multiple item types and underlying latent constructs, which subsumes previous work done by others. In addition, we provide corrections for multiple estimators of θ, the latent trait, including MLE, MAP and WLE. We provide analytical derivations that justifies our proposed correction, as well as simulation studies to examine the performance of the proposed correction with finite test lengths. An applied example is also provided to demonstrate proof of concept. We conclude with recommendations for practitioners when the asymptotic correction works well under different conditions and also future directions.

Key words: person fit, item response theory, l_z, Asymptotics, outlier detection, multidimensional, mixed item type.

1. Introduction

Person fit is an important topic within the item response theory (IRT) literature (Meijer, 1996; Meijer & Sijtsma, 2001. Person fit aims to classify individual response patterns as typical or atypical given an assumed measurement model (Meijer & Sijtsma, 2001). There has been much development in defining person fit statistics in the literature. For a recent review of applications and theoretical development of person fit statistics, see Rupp (2013). Person fit statistics have been utilized on both educational and psychological assessment data. For instance, person fit statistics have been used to identify aberrant (e.g. careless) responses on an survey or test (Conijn, Emons, & Sijtsma, 2014). Person fit has also been used to screen subtypes of suicide for clinicians by identifying individuals who do not demonstrate typical response patterns (Conrad et al., 2010).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09756-3.

Ying Cheng is supported by the National Science Foundation Grant SES-1853166. The contribution of Lizhen Lin was supported by NSF grant DMS Career 1654579.

Correspondence should be made to Ying Cheng, Department of Psychology, University of Notre Dame, 442 Corbett Family Hall, Notre Dame, IN46556, USA. Email: ycheng4@nd.edu

TABLE 1.
Summary of research on asymptotics of l_z.

	Dichotomous	Polytomous
Unidimensional	Bedrick (1997); Molenaar and Hoijtink (1990); Snijders (2001)	Sinharay (2016a); Tendeiro (2017); von Davier and Molenaar (2003)
Multidimensional	Albers et al. (2016)	None

We will focus on person fit statistics based on the l_z statistic (Drasgow, Levine, & Williams, 1985). There have been many studies on the l_z statistic and its applications. For instance, it has been shown that l_z is very powerful in detecting aberrant individuals when compared to other person fit statistics (Karabatsos, 2003). l_z has also been shown to preform well when used to identify multiple types of aberrant responses (Hong, Steedle, & Cheng, 2019; Niessen, Meijer, & Tendeiro, 2016). Moreover, l_z is a widely used statistic as it is offered in several software packages such as mirt (Chalmers, 2012), ltm (Rizopoulos, 2006) and Perfit (Tendeiro, Meijer, & Niessen, 2016).

There have been several studies examining statistical properties of the l_z statistic, with a focus on l_z's asymptotic properties (Albers, Meijer, & Tendeiro, 2016; Sinharay, 2016a; Snijders, 2001; Tendeiro, 2017). Through simulation studies, researchers found that l_z did not follow the theoretical properties outlined by Drasgow et al. (1985) in practice due to estimation of a latent trait. The consequences of this discrepancy is over classifying individuals as following the assumed measurement model and reduced power when detecting aberrant individuals (Reise, 1990). Several corrections have been proposed in the literature. For instance, Molenaar and Hoijtink (1990) proposed a Chi-squared approximation for the Rasch model. Subsequent extensions were proposed using alternative approximations, such as Edgeworth approximations, and for polytomous Rasch models and latent class models (Bedrick, 1997; von Davier & Molenaar, 2003). In the following, we will focus on the latent trait model framework and the correction proposed by Snijders (2001). Snijders (2001) rigorously investigated the asymptotic properties of the l_z statistic and proposed a corrected statistic known as l_z^* for dichotomous items. l_z^* was shown to perform better when classifying individuals by maintaining better nominal type I error rates under the two parameter logistic item response model. Several other researchers have built on Snijders (2001) original work by considering scales beyond dichotomous items (Sinharay, 2016a), various types of item response models (Tendeiro, 2017), and measuring multiple traits (Albers, Meijer, & Tendeiro, 2016). Table 1 summarizes previous studies that investigated the asymptotic properties of l_z and l_z^* under unidimensional and multidimensional models.

Table 1 shows an obvious gap in the development of the l_z^* statistic for applied research. Many researchers collect data with items of varying response categories and contains multiple subscales, measuring closely related, yet distinct, constructs. This is very common in psychological research. For instance, the Big Five inventory measures five latent constructs and can potentially have varying item types depending on which version is used (Goldberg & Kilkowski, 1985). In Albers et al. (2016), the authors develop an l_z^* statistic for scales with dichotomous items and multiple subscales. Their simulations provide evidence that l_z^* for multiple subscales may not asymptotically follow a standard normal distribution. More analytical work can be done to further investigate the properties of l_z^* when measuring multiple constructs. The properties of l_z^* have not been studied in some situations such as for a varying number of latent traits and number of categories within each item.

This paper aims to fill these gaps with theoretical work and simulation studies. The rest of the paper is organized as follows. First, we will review l_z for mixed-format scales and for multiple constructs. Next, we will review asymptotic corrections for both cases. We will then propose a more general framework for l_z^* that encompasses both a mixture of item types and multiple constructs. We will provide rigorous proofs for the newly proposed framework drawing from earlier work of Snijders (2001) and Sinharay (2016a). We will also provide a comprehensive simulation study. We will then provide an applied analysis to demonstrate the utility of the new l_z^* statistic and end with discussions and potential future directions.

2. Methods

2.1. Review of l_z for Mixed Item Types

The original l_z for mixed-format scales was proposed by Sinharay (2016a). Consider a respondent with a true ability, θ, that answers to p items that can be dichotomous or polytomous. Let

$$\mathbb{I}_j(X_i) = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where X_i is an item response, and $j = 0, \ldots, m_i$ with $m_i + 1$ representing the number of categories for item i. The log-likelihood of the examinee's scores can be defined as follows:

$$l = \sum_{i=1}^{p} \sum_{j=0}^{m_i} \mathbb{I}_j(X_i) \log P_{ij}(\theta), \tag{2}$$

where $P_{ij}(\theta)$ is the probability that the respondent with ability θ endorsed category j of item i. For the graded response model (GRM; (Samejima, 1969)) the probability of endorsing category j or above can be characterized by a 2-parameter logistic model (2-PL; (Birnbaum, 1968)):

$$P_{ij}(\theta)^* = \frac{1}{1 + \exp(-a_i(\theta - b_{ij}))}, \tag{3}$$

where b_{ij} is the location parameter representing the boundary that separates the $(j-1)^{th}$ and j^{th} response category of item i; a_i is the discrimination parameter for item i for all boundary functions of item i. The probability of endorsing response option j can be expressed by:

$$P_{ij}(\theta) = P_{ij}(\theta)^* - P_{i(j+1)}(\theta)^*. \tag{4}$$

The probability of responding below the first option and above the highest option is set to 0. Note if item i is dichotomous, then $m_i = 1$ and the GRM reduces to the 2-PL model. Equation (2) does not place any restrictions on the number of item response categories. The number of response categories can be different from one item to another. Equation (2) implies the following:

$$E(l|\theta) = \sum_{i=1}^{p} \sum_{j=0}^{m_i} P_{ij}(\theta) \log P_{ij}(\theta) \tag{5}$$

and

$$\text{Var}(l|\theta) = \sum_{i=1}^{p} \mathbf{u}_i'(\theta)\mathbf{D}_i(\theta)\mathbf{u}_i(\theta), \tag{6}$$

where

$$\mathbf{u}_i'(\theta) = (\log P_{i0}(\theta), \log P_{i1}(\theta), \ldots, \log P_{im_i}(\theta)), \tag{7}$$

and

$$\mathbf{D}_i(\theta) = \begin{bmatrix} P_{i0}(\theta)(1 - P_{i0}(\theta)) & -P_{i0}(\theta)P_{i1}(\theta) & \ldots & -P_{i0}(\theta)P_{im_i}(\theta) \\ -P_{i1}(\theta)P_{i0}(\theta) & P_{i1}(\theta)(1 - P_{i1}(\theta)) & \ldots & -P_{i1}(\theta)P_{im_i}(\theta) \\ \ldots & \ldots & \ldots & \ldots \\ -P_{im_i}(\theta)P_{i0}(\theta) & -P_{im_i}(\theta)P_{i1}(\theta) & \ldots & P_{im_i}(\theta)(1 - P_{im_i}(\theta)) \end{bmatrix}. \tag{8}$$

The l_z statistic for a mixed-format test is then defined as follows:

$$l_z = \frac{l - E(l|\theta)}{\sqrt{\text{Var}(l|\theta)}}, \tag{9}$$

which follows a standard normal distribution.

2.2. *Review of l_{zm} for Multidimensional Constructs*

Drasgow, Levine, and McLaughlin (1991) developed a statistic for multiple subtests with dichotomous items. The proposed statistic is general enough to encompass both dichotomous and polytomous items on a scale. For each subtest s, where $s = 1, \ldots, S$, Drasgow et al. (1991) defined l_z with multiple subscales, l_{zm}. Note that a scale with multiple subscales where each item measures a single latent trait is a scale that has simple structure (Zhang & Stout, 1999). Let l_s, $E(l_s)$ and $\text{Var}(l_s)$ be the log-likelihood function and its first two moments for a specific subtest. Then,

$$l_{zm} = \frac{\sum_{s=1}^{S}\left(l_s\right) - \sum_{s=1}^{S}\left(E(l_s)\right)}{\left(\sum_{s=1}^{S}\text{Var}(l_s)\right)^{1/2}}. \tag{10}$$

For notational purposes, we can rewrite the model as a multidimensional item response model for mixed-format scales with simple structure (Reckase, 2009; Zhang & Stout, 1999). The log-likelihood of a participant's scores can be defined as follows:

$$l = \sum_{i=1}^{p}\sum_{j=0}^{m_i} \mathbb{I}_j(X_i)\log P_{ij}(\boldsymbol{\theta}), \tag{11}$$

where X_i, the response on item i, is an integer between 0 and m_i. $P_{ij}(\boldsymbol{\theta})$ is the probability that an examinee with ability $\boldsymbol{\theta}$ endorsed category j on item i where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_S)'$ denotes a set

of latent traits. The probability of endorsing category j or above can be characterized by a 2-PL multidimensional item response model:

$$P_{ij}^{*}(\boldsymbol{\theta}) = \frac{1}{1 + \exp(-(\boldsymbol{a}_i'\boldsymbol{\theta} - b_{ij}))}, \tag{12}$$

where $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iS})'$ is a vector of S discrimination parameters reflecting the relative importance of each dimension when answering an item correctly. Here, b_{ij} is the location parameter for the boundary that separates the $(j-1)^{th}$ and j^{th} response category of item i. The probability of endorsing response option j can be expressed by:

$$P_{ij}(\boldsymbol{\theta}) = P_{ij}(\boldsymbol{\theta})^{*} - P_{i(j+1)}(\boldsymbol{\theta})^{*}. \tag{13}$$

Equation (11) implies the following:

$$E(l|\boldsymbol{\theta}) = \sum_{i=1}^{p}\sum_{j=0}^{m_i} P_{ij}(\boldsymbol{\theta}) \log P_{ij}(\boldsymbol{\theta}) \tag{14}$$

and

$$\mathrm{Var}(l|\boldsymbol{\theta}) = \sum_{i=1}^{p} \mathbf{u}_{\mathbf{i}}'(\boldsymbol{\theta})\mathbf{D}_{\mathbf{i}}(\boldsymbol{\theta})\mathbf{u}_{\mathbf{i}}(\boldsymbol{\theta}), \tag{15}$$

where

$$\mathbf{u}_i'(\boldsymbol{\theta}) = (\log P_{i0}(\boldsymbol{\theta}), \log P_{i1}(\boldsymbol{\theta}), \ldots, \log P_{im_i}(\boldsymbol{\theta})) \tag{16}$$

and

$$\mathbf{D}_{\mathbf{i}}(\boldsymbol{\theta}) = \begin{bmatrix} P_{i0}(\boldsymbol{\theta})(1 - P_{i0}(\boldsymbol{\theta})) & -P_{i0}(\boldsymbol{\theta})P_{i1}(\boldsymbol{\theta}) & \ldots & -P_{i0}(\boldsymbol{\theta})P_{im_i}(\boldsymbol{\theta}) \\ -P_{i1}(\boldsymbol{\theta})P_{i0}(\boldsymbol{\theta}) & P_{i1}(\boldsymbol{\theta})(1 - P_{i1}(\boldsymbol{\theta})) & \ldots & -P_{i1}(\boldsymbol{\theta})P_{im_i}(\boldsymbol{\theta}) \\ \ldots & \ldots & \ldots & \ldots \\ -P_{im_i}(\boldsymbol{\theta})P_{i0}(\boldsymbol{\theta}) & -P_{im_i}(\boldsymbol{\theta})P_{i1}(\boldsymbol{\theta}) & \ldots & P_{im_i}(\boldsymbol{\theta})(1 - P_{im_i}(\boldsymbol{\theta})) \end{bmatrix}. \tag{17}$$

Similar to the unidimensional case, one can rewrite a multidimensional l_z for multidimensional test, l_{zm}, as

$$l_{zm} = \frac{l - E(l|\boldsymbol{\theta})}{\sqrt{\mathrm{Var}(l|\boldsymbol{\theta})}}. \tag{18}$$

It is important to note that the variance term in Equation (18) assumes local independence, $cov(X_i, X_{i'}|\boldsymbol{\theta}) = 0$ for all $i \neq i'$, which is a typical assumption for item response models. The latent traits do not have to be orthogonal. In the following sections, we review the asymptotic corrections for l_z and l_{zm}.

2.3. Review of l_z^* for Mixed Item Types

Sinharay (2016a) proposed that l_z can be expressed as a case of broader person fit statistics of the following form:

$$\frac{W(\theta)}{\sqrt{\text{Var}(W(\theta))}} \tag{19}$$

where,

$$W(\theta) = \sum_{i=1}^{p} \sum_{j=0}^{m_i} [\mathbb{I}_j(X_i) - P_{ij}(\theta)] w_{ij}(\theta), \tag{20}$$

an appropriate weight function is $w_{ij}(\theta)$. A suitable weight function means a weight function that corresponds to an identified person fit statistic (Magis, Raîche, & Béland, 2012). For a scale with mixed item types, the weight function is

$$w_{ij}(\theta) = \log P_{ij}(\theta). \tag{21}$$

Equation (21) implies the following:

$$E(W(\theta)) = 0 \text{ and } Var(W(\theta)) = p\sigma^2(\theta) \tag{22}$$

where,

$$\sigma^2(\theta) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{u_i'}(\theta) \mathbf{D_i}(\theta) \mathbf{u_i}(\theta). \tag{23}$$

A suitable estimator for θ such as the maximum likelihood estimate (MLE; $\hat{\theta}_{ML}$), maximum a posteriori (MAP; $\hat{\theta}_{MAP}$), or Warm (1989)'s weighted likelihood estimator (WLE; $\hat{\theta}_{WLE}$) can be plugged into the following equation:

$$l_z(\hat{\theta}) = \frac{W(\hat{\theta})}{\sqrt{p}\sigma(\hat{\theta})}. \tag{24}$$

Sinharay (2016a) used a first-order Taylor series approximation to express $l_z(\hat{\theta})$

$$\begin{aligned} \frac{1}{\sqrt{p}} W(\hat{\theta}) \approx \frac{1}{\sqrt{p}} W(\theta) + \sqrt{p}(\hat{\theta} - \theta) \Bigg\{ \frac{1}{p} \sum_{i=1}^{p} \Bigg(\sum_{j=0}^{m_i} (\mathbb{I}_j(X_i) - P_{ij}(\theta)) w'_{ij}(\theta) \Bigg) \\ - \frac{1}{p} \sum_{i=1}^{p} \Bigg(\sum_{j=0}^{m_i} P'_{ij}(\theta) w_{ij}(\theta) \Bigg) \Bigg\}. \end{aligned} \tag{25}$$

Note that $P'_{ij} = \frac{\partial P_{ij}}{\partial \theta}$ and $w'_{ij} = \frac{\partial w_{ij}}{\partial \theta}$. Sinharay (2016a) posited that the term $\frac{1}{p} \sum_{i=1}^{p} \left(\sum_{j=0}^{m_i} P'_{ij}(\theta) w_{ij}(\theta) \right)$ will not tend toward zero as the number of items increases. Therefore, one can replace $w_{ij}(\theta)$ with $\tilde{w}_{ij}(\theta)$ which satisfies the condition, $\sum_{i=1}^{p} \sum_{j=0}^{m_i} P'_{ij}(\theta) \tilde{w}_{ij}(\theta) = 0$. Sinharay (2016a) chose $\hat{\theta}$ that satisfies the condition,

$$r_0(\hat{\theta}) + \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big[\mathbb{I}_j(X_i) - P_{ij}(\hat{\theta})\Big]r_{ij}(\hat{\theta}) = 0 \text{ for some } r_0(\hat{\theta}) \text{ and } r_{ij}(\hat{\theta}) \tag{26}$$

If one uses $\hat{\theta}_{ML}$, then $r_0(\hat{\theta}) = 0$ and $r_{ij}(\hat{\theta}) = \frac{P'_{ij}(\theta)}{P_{ij}(\theta)}$ (see Sinharay (2016a) for other estimators). Let $\tilde{w}_{ij}(\hat{\theta}) = w_{ij}(\hat{\theta}) - c_n(\hat{\theta})r_{ij}(\hat{\theta})$ where

$$c_n(\hat{\theta}) = \frac{\sum_i \sum_{j=0}^{m_i} P'_{ij}(\hat{\theta})w_{ij}(\hat{\theta})}{\sum_i \sum_{j=0}^{m_i} P'_{ij}(\hat{\theta})r_{ij}(\hat{\theta})}. \tag{27}$$

Then,

$$l_z^*(\hat{\theta}) = \frac{W(\hat{\theta}) + c_n(\hat{\theta})r_0(\hat{\theta})}{\sqrt{p}\tau(\hat{\theta})}, \tag{28}$$

where $\tau^2(\hat{\theta}) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{v}'_i(\theta)\mathbf{D}_i(\theta)\mathbf{v}_i(\theta)$. $\mathbf{D}_i(\theta)$ is defined in Equation (17) and

$$\mathbf{v}'_i(\theta) = (\log \tilde{w}_{i0}(\theta), \log \tilde{w}_{i1}(\theta), \ldots, \log \tilde{w}_{im_i}(\theta)) \tag{29}$$

Equation (28) is derived by adding $c_n(\hat{\theta})r_0(\hat{\theta})$ in the numerator and replacing $\sigma^2(\hat{\theta})$ with $\tau^2(\hat{\theta})$ in Equation (24). Sinharay (2016a) demonstrated that l_z^* with mixed item types is justified to follow a standard normal distribution with the correction through a series of derivations and simulations.

2.4. Proposed l_{zm}^* for Multidimensional Constructs and Mixed Item Types

Albers et al. (2016) proposed a correction for person fit statistics when the scales measure multiple constructs with dichotomous items, l_{zm}^*. The authors used heuristics based on Snijders (2001) to justify why l_{zm}^* should follow a standard normal distribution. However, Albers et al. (2016) found in their simulation study that l_{zm}^* does not achieve a 0.05 type I error as expected across all of their simulation conditions. The researchers justified a discrepancy between their analytical work and simulations with the following arguments. First, they were limited by the number of simulation replications and conditions. Second, the approximation is only asymptotic and the sample size is finite. It is important to note that there are problems with those justifications. The asymptotic properties of l_z^* and l_{zm}^* are asymptotic with respect to the number of items, not the number of persons, as implied by Albers et al. (2016).

There are also limitations to the simulations in Albers et al. (2016) beyond the number of simulation and replication conditions. First, the performance of l_{zm}^* should differ as a function α levels such as 0.01 or 0.1. Albers et al. (2016) used only a cut-off with a corresponding α level of 0.05. Snijders (2001) and Sinharay (2016a) found that depending on the critical value, sometimes the asymptotic correction over or under correct the null distribution. Second, the number of dimensions may impact l_{zm}^*. Albers et al. (2016) only focused on scales with four latent traits. Varying the number of latent traits may also impact the performance of l_{zm}^*. Finally, Albers et al. (2016) was restricted only to the dichotomous case. Oftentimes, scales contain a mixture of item types.

To address these limitations from previous works, we provide more general proofs in order to investigate the asymptotic properties of l_{zm}^* for multidimensional constructs and mixed format.

Our study builds upon Albers et al. (2016) by providing rigorous proofs, extensions to mixed item types, and extensive simulation study.

The general form of the l_{zm} statistic can be written in the following form:

$$l_{zm} = \frac{M(\boldsymbol{\theta})}{\sqrt{\text{Var}(M(\boldsymbol{\theta}))}}, \tag{30}$$

where

$$M(\boldsymbol{\theta}) = \sum_{i=1}^{p} \sum_{j=0}^{m_i} (\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}))w_{ij}(\boldsymbol{\theta}). \tag{31}$$

Similar to the unidimensional case, one can define $w_{ij}(\boldsymbol{\theta})$ as:

$$w_{ij}(\boldsymbol{\theta}) = \log P_{ij}(\boldsymbol{\theta}). \tag{32}$$

Equation (31) implies the following:

$$E(M(\boldsymbol{\theta})) = 0 \text{ and } \text{Var}(M(\boldsymbol{\theta})) = \gamma^2(\boldsymbol{\theta}), \tag{33}$$

where

$$\gamma^2(\boldsymbol{\theta}) = \sum_{i=1}^{p} \mathbf{u}_i'(\boldsymbol{\theta})\mathbf{D}_i\mathbf{u}_i(\boldsymbol{\theta}). \tag{34}$$

Therefore, $l_{zm}(\boldsymbol{\theta})$ is:

$$l_{zm}(\boldsymbol{\theta}) = \frac{M(\boldsymbol{\theta})}{\sqrt{\text{Var}(M(\boldsymbol{\theta}))}} = \frac{\sum_{i=1}^{p} \sum_{j=0}^{m_i} (\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}))w_{ij}(\boldsymbol{\theta})}{\sqrt{\sum_{i=1}^{p} \mathbf{u}_i'(\boldsymbol{\theta})\mathbf{D}_i\mathbf{u}_i(\boldsymbol{\theta})}}. \tag{35}$$

Following notation by Magnus and Neudecker (1988), we can approximate $\frac{1}{\sqrt{p}}M(\hat{\boldsymbol{\theta}})$ with a first-order Taylor series approximation:

$$\frac{1}{\sqrt{p}}M(\hat{\boldsymbol{\theta}}) \approx \frac{1}{\sqrt{p}}M(\boldsymbol{\theta}) + \sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\Big(\frac{1}{p}\nabla M(\boldsymbol{\theta})\Big), \tag{36}$$

where $\nabla M(\boldsymbol{\theta})$ is a $S \times 1$ gradient vector of $M(\boldsymbol{\theta})$:

$$\nabla M(\boldsymbol{\theta}) = \sum_{i=1}^{p} \Big\{ \sum_{j=0}^{m_i} (\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}))\Big(\nabla w_{ij}(\boldsymbol{\theta})\Big)\Big\} - \sum_{i=1}^{p} \Big\{ \sum_{j=0}^{m_i} \Big(\nabla P_{ij}(\boldsymbol{\theta})\Big)w_{ij}(\boldsymbol{\theta})\Big\}. \tag{37}$$

The term $\frac{1}{\sqrt{p}}M(\boldsymbol{\theta})$ has an asymptotically normal distribution. Note that $\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ in the second term is bounded around a neighborhood of the true latent trait, $\boldsymbol{\theta}$. Among the two averages in

Equation (37), the first is a vector of bounded random variables with an expectation of $\mathbf{0}$ as the number of items increases. However, the second term is a vector that will not tend toward $\mathbf{0}$ with an increasing number of items. Thus, the asymptotic null distribution will not follow a standard normal distribution. In the following section, we demonstrate how one can correct the null distribution. Suppose $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ satisfies the following condition:

$$\begin{bmatrix} t_{01}(\boldsymbol{\theta}) \\ t_{02}(\boldsymbol{\theta}) \\ \vdots \\ t_{0S}(\boldsymbol{\theta}) \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij1}(\boldsymbol{\theta}) \\ \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij2}(\boldsymbol{\theta}) \\ \vdots \\ \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ijS}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{38}$$

which can be rewritten as:

$$t_0(\boldsymbol{\theta}) + \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij}(\boldsymbol{\theta}) = \mathbf{0} \tag{39}$$

for some functions $t_0(\boldsymbol{\theta}) = (t_{01}, t_{02}, \ldots, t_{0S})'$ and $t_{ij}(\boldsymbol{\theta}) = (t_{ij1}, t_{ij2}, \ldots, t_{ijS})'$. For instance, $\hat{\boldsymbol{\theta}}_{ML}$ is the value of $\boldsymbol{\theta}$ for which:

$$\nabla l(X|\boldsymbol{\theta}) = \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)P_{ij}^{-1}(\boldsymbol{\theta})\nabla P_{ij}(\boldsymbol{\theta}) = \mathbf{0}. \tag{40}$$

The equality in Equation (39) holds where $t_0(\boldsymbol{\theta})$ and $t_{ij}(\boldsymbol{\theta})$ satisfy:

$$t_0(\boldsymbol{\theta}) = \mathbf{0} \text{ and } t_{ij}(\boldsymbol{\theta}) = P_{ij}^{-1}(\boldsymbol{\theta})\nabla P_{ij}(\boldsymbol{\theta}). \tag{41}$$

Expressions for other estimators are provided in Appendix A. We propose to replace $w_{ij}(\boldsymbol{\theta})$ with $\tilde{w}_{ij}(\boldsymbol{\theta})$ where each second term in Equation (37) follows $\sum_{i=1}^{p}\sum_{j=1}^{m_i}\nabla P_{ij}(\boldsymbol{\theta})\tilde{w}_{ij}(\boldsymbol{\theta}) = \mathbf{0}$. The new weights can be defined as:

$$\tilde{w}_{ij}(\boldsymbol{\theta}) = w_{ij}(\boldsymbol{\theta}) - c'(\boldsymbol{\theta})t_{ij}(\boldsymbol{\theta}) \tag{42}$$

where $t_{ij}(\boldsymbol{\theta})$ is a $S \times 1$ row vector for each item i and category j as defined in Equation (39). Let $c(\boldsymbol{\theta})$ be an $S \times 1$ vector where:

$$c(\boldsymbol{\theta}) = \Big(\sum_{i=1}^{p}\sum_{j=0}^{m_i}(\nabla P_{ij}(\boldsymbol{\theta}))'t_{ij}(\boldsymbol{\theta})\Big)^{-1}\Big(\sum_{i=1}^{p}\sum_{j=0}^{m_i}w_{ij}(\boldsymbol{\theta})\nabla P_{ij}(\boldsymbol{\theta})\Big). \tag{43}$$

Then,

$$l_{zm}^{*}(\boldsymbol{\theta}) = \frac{M(\boldsymbol{\theta}) + c'(\boldsymbol{\theta})t_0(\boldsymbol{\theta})}{\sqrt{p}\gamma(\boldsymbol{\theta})}, \tag{44}$$

where

$$\gamma^2(\boldsymbol{\theta}) = \sum_{i=1}^{p} \mathbf{v}_i'(\boldsymbol{\theta}) \mathbf{D}_i \mathbf{v}_i(\boldsymbol{\theta}), \tag{45}$$

$\mathbf{D}_i$ is defined in Equation (17), $t_0(\boldsymbol{\theta})$ is defined in Equation (39), and

$$\mathbf{v}_i(\boldsymbol{\theta}) = (\tilde{w}_{i0}(\boldsymbol{\theta}), \tilde{w}_{i1}(\boldsymbol{\theta}), \ldots, \tilde{w}_{im_i}(\boldsymbol{\theta})). \tag{46}$$

If there is only one latent trait, $S = 1$, then l_{zm}^* defined in Equation (44) reduces to l_z^* defined by Sinharay (2016a) in Equation (28). Similarly, if there are only dichotomous items, $m_j = 1 \; \forall \; j = 1, \ldots, p$, then l_{zm}^* defined in Equation (44) reduces to l_z^* defined by Albers et al. (2016). Recall that Albers et al. (2016) found that the dichotomous version of l_{zm}^* did not follow its asymptotic null distribution in practice for dichotomous items. In the following section, we provide a set of proofs of the properties of l_{zm}^* using similar arguments made by Sinharay (2016a) and Snijders (2001) to further investigate this phenomenon.

Consider a sequence of item responses $X_1, \ldots, X_p$ from a mixed format test given in Equations (12) and (13). Denote $\boldsymbol{\theta}_0$ as the true parameter value for the clarity of notation.

Theorem 1. *Under the following assumptions:*

1. $\hat{\boldsymbol{\theta}}$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*, and* $\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *has an asymptotic nondegenerate distribution.*
2. $P_{ij}(\boldsymbol{\theta})$ *and* $w_{ij}(\boldsymbol{\theta})$ *are twice differentiable and uniformly bounded over a small neighborhood of* $\boldsymbol{\theta}_0$*. Their first- and second-order derivatives are also uniformly bounded in a neighborhood of* $\boldsymbol{\theta}_0$*.*

One has:

$$\frac{1}{\sqrt{p}}\left(\tilde{M}(\hat{\boldsymbol{\theta}}) - \tilde{M}(\boldsymbol{\theta}_0)\right) \to 0 \textit{ in probability}. \tag{47}$$

Proof. Note that

$$\begin{aligned} p^{-1/2}(\tilde{M}(\hat{\boldsymbol{\theta}}) - \tilde{M}(\boldsymbol{\theta}_0)) =& p^{-1/2} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \left(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\right)\left(\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) - \tilde{w}_{ij}(\boldsymbol{\theta}_0)\right) \\ & - p^{-1/2} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \left(P_{ij}(\hat{\boldsymbol{\theta}}) - P_{ij}(\boldsymbol{\theta}_0)\right)\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}). \end{aligned} \tag{48}$$

We will show that each of the above two terms converges to zero in probability as $p \to \infty$. The first term on the right-hand side can be rewritten as:

$$\begin{aligned} & p^{-1/2} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big(\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) - \tilde{w}_{ij}(\boldsymbol{\theta}_0)\Big) \\ & = \sqrt{p}\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big(\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) - \tilde{w}_{ij}(\boldsymbol{\theta}_0)\Big)\Big) \\ & = \sqrt{p}\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\Big(\nabla \tilde{w}_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}\Big), \end{aligned} \tag{49}$$

where $\nabla \tilde{w}_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}$ is a $S \times 1$ gradient vector of $\tilde{w}_{ij}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^*$ which lies in the line segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. The above equation can be rewritten as:

$$= \sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big(\nabla \tilde{w}_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}\Big). \tag{50}$$

By law of large numbers (see section 6.2 in Bhattacharya, Lin, and Victor (2016)), one has

$$\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big(\nabla \tilde{w}_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}\Big) \to 0 \text{ in probability.} \tag{51}$$

By assumption 1, $\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to some random vector with a nondegenerate distribution. Then by Slutsky's lemma (see section 5.5 in Casella and Berger (2001)).

$$\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\Big(\nabla \tilde{w}_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}\Big)\Big) \to 0 \text{ in probability.} \tag{52}$$

Now we proceed to show that the second term in Equation (48) also converges to zero in probability.

$$\begin{aligned} & p^{-1/2} \sum_{i=1}^{p} \sum_{j=1}^{m_i} (P_{ij}(\hat{\boldsymbol{\theta}}) - P_{ij}(\boldsymbol{\theta}_0))\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) \\ & = \sqrt{p}\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\nabla P_{ij}(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}} + o(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||)\Big)\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) \\ & = \sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\Big(\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \Big(\nabla P_{ij}(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}} + o(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||\Big)\tilde{w}_{ij}(\hat{\boldsymbol{\theta}})\Big). \end{aligned} \tag{53}$$

Note that $\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{m_i}\nabla P_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, and $\nabla P_{ij}(\boldsymbol{\theta})$ and $\tilde{w}_{ij}(\boldsymbol{\theta})$ are all continuous and bounded functions in a neighborhood of $\boldsymbol{\theta}_0$. Then,

$$\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{m_i}\nabla\tilde{P}_{ij}(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}}\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) \to 0 \text{ in probability.} \tag{54}$$

Again, using the fact that $\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to some random vector in distribution and by Slutsky's theorem,

$$\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big(\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{m_i}\nabla\tilde{P}_{ij}(\hat{\boldsymbol{\theta}}) + o(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||)\Big)\tilde{w}_{ij}(\hat{\boldsymbol{\theta}}) \to 0. \tag{55}$$

Note that the remaining term $\sqrt{p}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big(\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{m_i} o(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||)\tilde{w}_{ij}(\hat{\boldsymbol{\theta}})\Big) \to 0$ is negligible, due to the fact that the remainder term vanishes as $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| \to 0$ as $p \to \infty$ and $\tilde{w}_{ij}(\hat{\boldsymbol{\theta}})$ are uniformly bounded. □

Theorem 2. *Assume the second assumption in Theorem 1 holds and $\gamma^2(\boldsymbol{\theta}_0) < \infty$ as $p \to \infty$. Then one has*

$$\frac{1}{\sqrt{p}\gamma(\boldsymbol{\theta}_0)}\tilde{M}(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{p}\gamma(\boldsymbol{\theta}_0)}\sum_{i=1}^{p}\sum_{j=1}^{m_i}(\mathbb{I}_j(x_i) - P_{ij}(\boldsymbol{\theta}_0))\tilde{w}_{ij}(\boldsymbol{\theta}_0) \to N(0,1). \tag{56}$$

Proof. Let $z_i = \sum_{j=1}^{m_i}\Big(\mathbb{I}_j(x_i) - P_{ij}(\boldsymbol{\theta}_0)\Big)\tilde{w}_{ij}(\boldsymbol{\theta}_0)$. One has $E(z_i) = 0$, $\mathrm{Var}(z_i) = \mathbf{v}_\mathbf{i}'(\boldsymbol{\theta}_0)\mathbf{D}_\mathbf{i}(\boldsymbol{\theta}_0)\mathbf{v}_\mathbf{i}(\boldsymbol{\theta}_0)$ and $\mathrm{Var}(\sum_{i=1}^{p} z_i) = p\gamma^2(\boldsymbol{\theta}_0)$. Note that:

$$\max_{1\le i\le n}\left(\frac{\mathrm{Var}(z_i)}{\mathrm{Var}(\sum_{i=1}^{p} z_i)}\right) = \max_{1\le i\le n}\left(\frac{\mathbf{v}_\mathbf{i}'(\boldsymbol{\theta}_0)\mathbf{D}_\mathbf{i}(\boldsymbol{\theta}_0)\mathbf{v}_\mathbf{i}(\boldsymbol{\theta}_0)}{\sum_{i=1}^{p}\mathbf{v}_\mathbf{i}'(\boldsymbol{\theta}_0)\mathbf{D}_\mathbf{i}(\boldsymbol{\theta}_0)\mathbf{v}_\mathbf{i}(\boldsymbol{\theta}_0)}\right) \to 0 \tag{57}$$

since $\mathbf{v}_\mathbf{i}'(\boldsymbol{\theta}_0)\mathbf{D}_\mathbf{i}(\boldsymbol{\theta}_0)\mathbf{v}_\mathbf{i}(\boldsymbol{\theta}_0)$ are uniformly bound $\forall i = 1, \ldots, p$. Then by the Linderberg Central Limit Theorem (see appendix D in n Bhattacharya et al. (2016)),

$$\frac{\sum_{i=1}^{p} z_i - 0}{\sqrt{\mathrm{Var}(\sum_{i=1}^{p} z_i)}} = \frac{\tilde{M}(\boldsymbol{\theta}_0)}{\sqrt{p}\gamma(\boldsymbol{\theta}_0)} \to \mathrm{N}(0,1). \tag{58}$$

□

Theorem 3. *Assume assumption 2 holds, $\gamma^2(\boldsymbol{\theta}_0) < \infty$ as $p \to \infty$, and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$. Then, one has:*

$$\frac{1}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})}M(\hat{\boldsymbol{\theta}}) \to N(0,1) \tag{59}$$

Proof. Given that $M(\hat{\boldsymbol{\theta}}) = \tilde{M}(\hat{\boldsymbol{\theta}}) - c(\hat{\boldsymbol{\theta}})t_{ij}(\hat{\boldsymbol{\theta}})$, it implies,

$$\frac{1}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})}M(\hat{\boldsymbol{\theta}}) = \frac{\tilde{M}(\hat{\boldsymbol{\theta}})}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})} + \frac{\tilde{M}(\hat{\boldsymbol{\theta}}) - \tilde{M}(\boldsymbol{\theta}_0)}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})} - \frac{c'(\hat{\boldsymbol{\theta}})t_0(\hat{\boldsymbol{\theta}})}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})}. \tag{60}$$

Note that the second term on the right-hand side goes to zero since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, which implies $\tilde{M}(\hat{\boldsymbol{\theta}}) \to \tilde{M}(\boldsymbol{\theta}_0)$ since $\tilde{M}(\cdot)$ is continuous. The third term on the right-hand side also converges to zero, due to the uniform boundedness of $c(\hat{\boldsymbol{\theta}})$ and $t_0(\hat{\boldsymbol{\theta}})$ over a neighborhood of $\boldsymbol{\theta}_0$ ($\hat{\boldsymbol{\theta}}$ is in a small neighborhood of $\boldsymbol{\theta}_0$). Note that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$, which implies $M(\hat{\boldsymbol{\theta}}) \to M(\boldsymbol{\theta}_0)$ and $\gamma(\hat{\boldsymbol{\theta}}) \to \gamma(\boldsymbol{\theta}_0)$, and $\frac{\tilde{M}(\hat{\boldsymbol{\theta}})}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})} \to \frac{\tilde{M}(\boldsymbol{\theta}_0)}{\sqrt{p}\gamma(\boldsymbol{\theta}_0)}$ and the later converges in law to a standard normal distribution, by Slutsky's lemma

$$\frac{\tilde{M}(\hat{\boldsymbol{\theta}})}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})} \to N(0, 1), \tag{61}$$

Since the last two terms of Equation (60) converge to zero in probability. Therefore, l^*_{zm} can be written in the following form

$$l^*_{zm} = \frac{\tilde{M}(\hat{\boldsymbol{\theta}})}{\sqrt{p}\gamma(\hat{\boldsymbol{\theta}})}. \tag{62}$$

□

The above theorems have implications based on past research. First, the current proofs demonstrate that l^*_{zm} should asymptotically converge to a standard normal distribution with enough items, even with multiple traits, which supports the preliminary work done by Albers et al. (2016). Similarly, l^*_{zm} extends the derivations done by Sinharay (2016a) to encompass multiple latent traits. l^*_{zm} is able to accommodate mixed (dichotomous and polytomous) item types.

However, there remains several open questions. Albers et al. (2016) found empirical evidence with simulation studies that when there are multiple traits, l^*_{zm} does not achieve adequate type I error rates with dichotomous items. Furthermore, Sinharay (2016a) did not consider the impact of varying the number of item categories within each polytomous item. Tendeiro (2017) found that the number of item categories impact l^*_z for unidimensional cases. Moreover, there is no research on the impact of both the number of categories and latent traits and their impact on l^*_{zm}. The following simulation aims to fill in these gaps.

3. Simulation

A comprehensive simulation study was performed to examine the type I error rate and the power of l^*_{zm}. Two simulations were conducted in order to address both questions.

3.1. Simulation: Type I Error

Type I error is defined as the proportion of participants who conform to the assumed measurement model that are identified as aberrant responses. In the current simulation, the true model is a multidimensional graded response model. Previous work found little differences between using different estimators (Sinharay, 2016a). For latent trait estimation, we used maximum likelihood

estimation with bounds between -3 to 3. Item parameters are assumed to be known, which is a typical assumption in the person fit literature. Responses were simulated from a multidimensional graded response model with 2 to 5 response categories that have simple structure. Item parameters were generated similar to Hong et al. (2019) where the discrimination parameter was sampled from a uniform distribution with a lower bound of .5 and upper bound of 2. The location parameters were sampled with equal distances from a range with lower bound of -1.5 and upper bound of 1.5 and perturbed by a random variable drawn from a uniform distribution ranging from -0.3 to 0.3. When there were two categories, we randomly sampled the location parameters from -1.8 to 1.8. We also varied the number of latent traits ranging from 1 to 5. For each dimension, we generated latent traits drawn from -2, -1, 0, 1, and 2. It is important to note that the sample size is not relevant because item parameters are assumed to be known. However, the number of items can influence the performance of l^*_{zm}. We varied the number of items per trait to be 16, 32, or 64.

In sum, we had a total of 300 simulation conditions for data generation with 100,000 replications for each condition, with four factors fully crossed (three scale lengths per latent trait, five number of latent traits, four number of categories, and five values of θ_s).

In each simulation condition, we evaluated the type I error rates with varying α levels: .01, .05, and .10 using either l_{zm} or l^*_{zm}.

3.2. Results: Type I Error

Tables 2, 3, and 4 present the results type I error rates when there are 16, 32, or 64 items per latent trait, respectively. Due to the large number of simulation results, we present results when $\theta = 0$ due to negligible impact of different latent trait levels on type I error rates. Results of other θ values are available upon request. In general, l_{zm} consistently achieved too conservative type I error rates across simulation conditions. Using an α level of 0.01 lead to liberal estimates using l^*_{zm}. On the other hand, using an α level of 0.1 lead to conservative estimates using l^*_{zm}, in general. Oftentimes researchers use α levels of 0.05. Therefore, we focus our discussion on these conditions.

There are a few takeaways from the tables. When an item has more categories, the type I error rate becomes more conservative. For instance, when there are two categories for a single trait with 16 items, the type I error rate is 0.058. When the number of categories increases to five, the type I error rate decreases to 0.043.

When the scale measures more than one latent traits (or consists of multiple subscales), the type I error rate becomes more conservative. For instance, when there is a single latent trait for dichotomous items, the type I error rate is 0.058. When the number of latent traits increases to five where each scale is measured by 16 items, the type I error rate decreases to 0.033.

If we fix the number of items on a test and increase the number of latent traits the type I error suffers. For instance, when there are 64 items in total that measure a single latent trait with two categories, the type I error rate is 0.052. If the same 64 items are split between two latent traits (32 items per trait), the type I error is 0.046.

With shorter scales, the impact of number of categories and latent traits becomes more evident. However, as the number of items increase per latent trait, the impact reduces. When there are three latent traits each measured by 16 items with three categories, the type I error rate is 0.030. When the number of items increase to 64 per latent trait, the type I error rate is 0.036. Even with the current simulations with 64 items it was not enough items to achieve perfect type I error rates. However, it would be unrealistic to assume more than 64 items in one subscale.

In order to further understand the impact of the number of categories and latent traits on the null distribution of l^*_{zm} and l_{zm}. We plot the empirical distributions from the simulation of l^*_{zm} and l_{zm} when there are 64 items in a scale with varying number of dimensions and categories in Fig. 1.

TABLE 2.
Type I error rates for l_z and l_z^* under various conditions when the number of items is 16 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.001	0.011	0.009	0.058	0.023	0.110
2	2	0.000	0.013	0.001	0.057	0.003	0.106
2	3	0.000	0.009	0.000	0.041	0.000	0.080
2	4	0.000	0.009	0.000	0.037	0.000	0.074
2	5	0.000	0.007	0.000	0.033	0.000	0.064
3	1	0.011	0.012	0.039	0.040	0.069	0.071
3	2	0.009	0.010	0.034	0.036	0.061	0.064
3	3	0.004	0.007	0.021	0.030	0.044	0.056
3	4	0.005	0.009	0.024	0.033	0.049	0.062
3	5	0.004	0.008	0.019	0.031	0.042	0.060
4	1	0.008	0.011	0.037	0.043	0.073	0.081
4	2	0.002	0.010	0.014	0.042	0.036	0.078
4	3	0.002	0.008	0.012	0.034	0.031	0.067
4	4	0.000	0.008	0.007	0.035	0.021	0.070
4	5	0.002	0.006	0.012	0.027	0.033	0.056
5	1	0.008	0.010	0.039	0.043	0.077	0.083
5	2	0.003	0.008	0.020	0.037	0.048	0.073
5	3	0.002	0.008	0.017	0.035	0.041	0.071
5	4	0.000	0.008	0.003	0.038	0.013	0.075
5	5	0.000	0.007	0.005	0.032	0.017	0.066

Note: Categories = number of categories per item, S = number of latent traits.

TABLE 3.
Type I error rates for l_z and l_z^* under various conditions when the number of items is 32 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.003	0.012	0.019	0.054	0.044	0.104
2	2	0.000	0.010	0.000	0.046	0.000	0.089
2	3	0.000	0.010	0.000	0.045	0.000	0.088
2	4	0.000	0.010	0.000	0.049	0.000	0.096
2	5	0.000	0.011	0.000	0.052	0.000	0.100
3	1	0.011	0.011	0.041	0.042	0.075	0.077
3	2	0.010	0.011	0.038	0.039	0.074	0.075
3	3	0.007	0.008	0.032	0.033	0.061	0.063
3	4	0.004	0.010	0.022	0.039	0.048	0.075
3	5	0.005	0.007	0.025	0.030	0.054	0.062
4	1	0.012	0.013	0.046	0.048	0.085	0.087
4	2	0.007	0.010	0.032	0.041	0.066	0.078
4	3	0.000	0.010	0.005	0.044	0.018	0.086
4	4	0.002	0.008	0.017	0.037	0.043	0.072
4	5	0.001	0.008	0.012	0.036	0.033	0.074
5	1	0.009	0.010	0.042	0.045	0.084	0.088
5	2	0.008	0.009	0.038	0.040	0.075	0.078
5	3	0.001	0.009	0.013	0.041	0.036	0.081
5	4	0.001	0.007	0.010	0.037	0.031	0.075
5	5	0.000	0.008	0.005	0.038	0.019	0.076

Note: Categories = number of categories per item, S = number of latent traits.

TABLE 4.
Type I error rates for l_z and l_z^* under various conditions when the number of items is 64 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.004	0.011	0.024	0.052	0.056	0.102
2	2	0.000	0.010	0.000	0.048	0.000	0.094
2	3	0.000	0.010	0.000	0.049	0.000	0.097
2	4	0.000	0.010	0.000	0.047	0.000	0.093
2	5	0.000	0.009	0.000	0.047	0.000	0.094
3	1	0.011	0.011	0.043	0.043	0.081	0.082
3	2	0.010	0.010	0.040	0.042	0.079	0.081
3	3	0.008	0.008	0.035	0.036	0.070	0.071
3	4	0.006	0.007	0.032	0.033	0.065	0.066
3	5	0.003	0.009	0.021	0.042	0.050	0.080
4	1	0.012	0.012	0.047	0.048	0.089	0.090
4	2	0.010	0.010	0.044	0.045	0.084	0.085
4	3	0.001	0.009	0.014	0.043	0.039	0.084
4	4	0.001	0.010	0.009	0.043	0.028	0.085
4	5	0.002	0.008	0.019	0.039	0.048	0.076
5	1	0.010	0.010	0.045	0.046	0.088	0.088
5	2	0.005	0.010	0.030	0.044	0.065	0.086
5	3	0.003	0.009	0.024	0.042	0.056	0.084
5	4	0.000	0.009	0.004	0.043	0.016	0.087
5	5	0.000	0.009	0.005	0.042	0.019	0.085

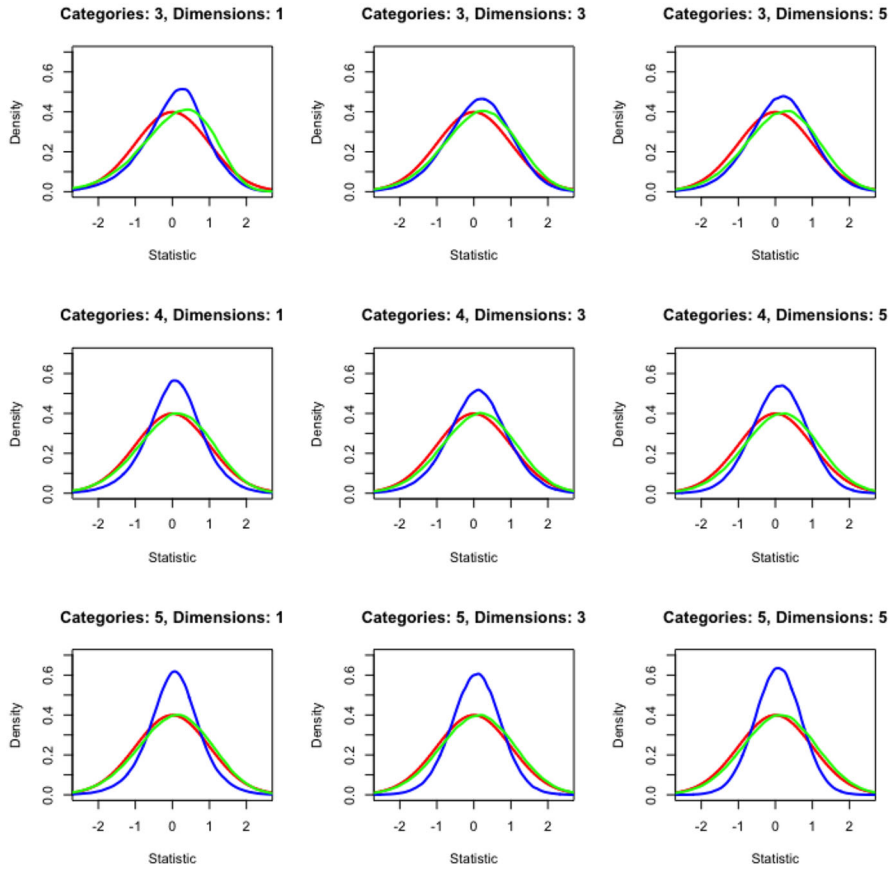
Note: Categories = number of categories per item, S = number of latent traits.

In general, l_{zm} shows a slight left skew when there were three response categories and a more peaked distribution with increasing number of response categories. For l_{zm}^*, the empirical distribution in general overlaps with the standard normal distribution compared to l_{zm}.

Our simulations demonstrate how increasing the number of items per scale improves the nominal type I error rate. However, increasing the number of latent traits, in general, reduces the type I error below nominal level with fixed or non-fixed total scale length. Moreover, increasing the number of categories within items also reduces the type I error. There is also an interaction between the manipulated factors, the scale length, the number of latent traits, and the number of item categories.

3.3. Simulation: Power

Power is defined as the proportion of flagged participants over the total number of participants affected by aberrant behavior. We follow the same data generation process as Hong et al. (2019) for aberrant responses. We generated both random and midpoint carelessness. Random carelessness was simulated where each response option has equal chance of endorsement. Midpoint carelessness was simulated where participants are more likely to select the middle categories. This was done by drawing from a Binomial distribution where the number of categories is equal to the number of item response options and the probability of success is fixed to .5. It is also important to consider the severity of careless responses when evaluating carelessness. Severity is the number of items in a response vector that does not conform to the item response model. In past research, others found that participants carelessly respond up to 50% of items on long surveys that involve multiple latent traits (Baer, Ballenger, Berru, & Wetter, 1997; Berry et al., 1992). We simulated the severity rate to be either 1/2 or 1/4 of the entire response vector.



Note: Red lines = N(0,1), Blue lines = l_z, and Green lines = l_z^*

FIGURE 1.
Distributions of l_z and l_z^* with varying number of categories and dimensions for 64 items. Note: Red lines = N(0,1), Blue lines = l_z, and Green lines = l_z^*. (Color figure online)

In sum, we had a total of 1200 simulation conditions for data generation with 100,000 replications for each condition, with six factors fully crossed (three scale lengths per latent trait, five number of latent traits, four number of categories, five values of θ_s, two levels of severity, and two types of aberrant responses).

3.4. Results: Power

Tables 5, 6, and 7 present the power to detect careless responses when there are 16, 32, or 64 items per latent trait, respectively. Due to the large number of simulation results, we focus our discussion when $\theta = 0$, severity is 1/2, and random responses. All other simulation conditions are available upon request. The power to detect individuals increased with increasing α levels and severity. For random and middle type responses, it was easier to detect the aberrance at more extreme latent traits (e.g., -2 or 2) compared to latent traits closer to the center of the distribution (e.g., 0). It is easier to detect random responses compared to middle type responses. In general, l_{zm} consistently achieved smaller power across simulation conditions when compared to l_{zm}^*. We focus our discussion when $\alpha = 0.05$.

TABLE 5.
Power for l_z and l_z^* under various conditions when the number of items is 16 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.096	0.113	0.233	0.254	0.343	0.363
2	2	0.248	0.326	0.465	0.530	0.592	0.640
2	3	0.680	0.726	0.845	0.869	0.907	0.921
2	4	0.745	0.786	0.893	0.911	0.940	0.950
2	5	0.966	0.980	0.991	0.994	0.996	0.997
3	1	0.375	0.389	0.576	0.586	0.681	0.689
3	2	0.408	0.417	0.619	0.626	0.725	0.729
3	3	0.743	0.772	0.878	0.891	0.924	0.932
3	4	0.965	0.975	0.988	0.991	0.994	0.995
3	5	0.958	0.964	0.988	0.989	0.994	0.995
4	1	0.120	0.138	0.280	0.302	0.398	0.417
4	2	0.284	0.375	0.521	0.597	0.658	0.712
4	3	0.770	0.833	0.897	0.926	0.940	0.956
4	4	0.409	0.441	0.631	0.653	0.740	0.754
4	5	0.704	0.776	0.877	0.907	0.934	0.949
5	1	0.126	0.147	0.292	0.316	0.410	0.432
5	2	0.211	0.342	0.437	0.555	0.577	0.667
5	3	0.354	0.462	0.608	0.685	0.736	0.788
5	4	0.629	0.722	0.828	0.876	0.901	0.927
5	5	0.586	0.775	0.822	0.908	0.906	0.950

Note: Categories = number of categories per item, S = number of latent traits.

TABLE 6.
Power for l_z and l_z^* under various conditions when the number of items is 32 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.244	0.260	0.447	0.461	0.568	0.579
2	2	0.664	0.711	0.851	0.875	0.916	0.928
2	3	0.990	0.993	0.997	0.998	0.999	0.999
2	4	0.961	0.966	0.990	0.991	0.995	0.996
2	5	0.974	0.981	0.994	0.996	0.997	0.998
3	1	0.747	0.752	0.883	0.886	0.930	0.932
3	2	0.986	0.988	0.995	0.996	0.997	0.998
3	3	0.985	0.986	0.996	0.997	0.998	0.998
3	4	1.000	1.000	1.000	1.000	1.000	1.000
3	5	1.000	1.000	1.000	1.000	1.000	1.000
4	1	0.328	0.337	0.553	0.561	0.672	0.678
4	2	0.648	0.669	0.824	0.835	0.891	0.898
4	3	0.909	0.929	0.970	0.976	0.985	0.988
4	4	0.955	0.957	0.989	0.989	0.995	0.995
4	5	0.896	0.904	0.967	0.969	0.985	0.986
5	1	0.194	0.201	0.407	0.414	0.543	0.549
5	2	0.412	0.450	0.640	0.666	0.751	0.769
5	3	0.740	0.851	0.911	0.950	0.958	0.974
5	4	0.740	0.804	0.902	0.927	0.950	0.962
5	5	0.909	0.965	0.980	0.992	0.993	0.997

Note: Categories = number of categories per item, S = number of latent traits.

TABLE 7.
Power for l_z and l_z^* under various conditions when the number of items is 64 per latent trait.

Categories	S	l_z : 1%	l_z^* : 1%	l_z : 5%	l_z^* : 5%	l_z : 10%	l_z^* : 10%
2	1	0.723	0.727	0.876	0.878	0.928	0.929
2	2	0.976	0.983	0.995	0.997	0.998	0.999
2	3	0.999	0.999	1.000	1.000	1.000	1.000
2	4	1.000	1.000	1.000	1.000	1.000	1.000
2	5	1.000	1.000	1.000	1.000	1.000	1.000
3	1	0.970	0.972	0.993	0.993	0.997	0.997
3	2	0.998	0.999	1.000	1.000	1.000	1.000
3	3	1.000	1.000	1.000	1.000	1.000	1.000
3	4	1.000	1.000	1.000	1.000	1.000	1.000
3	5	1.000	1.000	1.000	1.000	1.000	1.000
4	1	0.642	0.651	0.827	0.831	0.896	0.898
4	2	0.958	0.962	0.989	0.991	0.996	0.996
4	3	0.998	0.999	1.000	1.000	1.000	1.000
4	4	0.999	1.000	1.000	1.000	1.000	1.000
4	5	1.000	1.000	1.000	1.000	1.000	1.000
5	1	0.434	0.441	0.678	0.682	0.788	0.791
5	2	0.904	0.946	0.974	0.985	0.989	0.993
5	3	0.949	0.958	0.989	0.991	0.996	0.997
5	4	1.000	1.000	1.000	1.000	1.000	1.000
5	5	0.999	1.000	1.000	1.000	1.000	1.000

Note: Categories = number of categories per item, S = number of latent traits.

There are a few takeaways from the tables. When an item has more categories, the power to detect aberrant individuals decreased. For instance, when there are three categories for a single trait with 16 items per trait, the power is 0.586. When the number of categories increases to five, power decreased to 0.316. This pattern in general holds, except for when there are two categories. When we compare two categories to three categories for a single trait with 16 items, the power to detect aberrant individuals is 0.254 and 0.586, respectively. Moreover, there may be an within person effect size difference when we vary the number of categories within an item. For instance, an aberrant response when there are two categories (e.g., 0 changed to 1) has a larger standardized difference compared to five categories with the same change (e.g., 0 changed to 1 vs. 0 changed to 4). Further work should focus on developing a within person effect size measure.

The impact on the number of latent traits can be viewed in two ways. First, if a researcher is able to include more scales without having a finite test length, then the power to detect aberrant individuals increases. For instance, a scale with two categories measuring two latent traits with 16 items per latent trait (32 items total) has a power of 0.530. When a researcher includes another latent trait measured with 16 items (48 items total) the power increases to 0.869.

However, if a researcher has a finite test length, measuring more latent traits can increase, decrease, or not change the power to detect aberrant individuals. Consider a scale with four categories measuring two latent traits with 16 items per latent trait (32 items total) with a power of 0.597. If we measured only one latent trait, but keep the total number of items fixed at 32, then the power decreases to 0.561. On the other hand, a scale measuring two latent traits with 16 items with three categories per latent trait (32 items total) results in a power of 0.626. If we measured only one latent trait, but keep the total number of items fixed at 32, then the power increases to 0.866. Again, this finding is probably due to differences in within person effect sizes.

TABLE 8.
Detection Rates using the Big Five data set.

Statistic	Extraversion	Neuroticism	Agreeableness	Consciousness	Openness	Total
l_{zm}	0.073	0.075	0.077	0.076	0.096	0.125
l^*_{zm}	0.113	0.104	0.106	0.117	0.132	0.165

In general, when one increases the number of items on the scale, the power to detect aberrant individuals increase. These findings replicate previous research. Moreover, increasing the number of categories decreases the power to detect aberrant behavior. The impact of number of dimensions on power depends if the total scale length is fixed or not. If the total scale length is not fixed, then power increases with more latent traits (where the scale length increases with more items). With a finite scale length, power can increase, decrease, or not change depending on the number of item categories per latent trait. Our simulations have its limitations. The interpretation of aberrant behavior is confounded with varying item categories. The development of a within person effect size is necessary in order to compare our results, which is a potential area for further research.

4. Real Data Example

In order to illustrate the utility of l^*_{zm}, we apply it to data collected about the Big Five personality traits (Goldberg, 1992). Item response data were obtained from the Open Source Psychometrics Project (https://openpsychometrics.org). The respondents were from an online sample who consented to having their data stored and publicly available. The Big Five personality assessment consists five sub-scales, each with ten items, which measure five personality traits: extraversion, neuroticism, agreeableness, consciousness, and openness to experience. Developers of the platform found that the data collected on the website tend to be of good quality, at least relative to other platforms such as Amazon Mechanical Turk (https://openpsychometrics.org/rawdata/validity/). For instance, only 0.2% of respondents were flagged for content implausibility (such as reporting heights smaller than 4 feet). Other researchers have analyzed the Big 5 data, and found evidence supporting a five factor model (Jeon & De Boeck, 2019). However, Jeon and De Boeck (2019) suspected that subscales presented at the end of the assessment may contain fatigue effects, which interferes with content validity for the entire survey. Person fit analysis may help identify suspect individuals with fatigue effect or other aberrant responses.

In the current sample, 19,717 individuals responded to the survey with complete item responses. We split the data into two parts. The rationale for data splitting is to remove the confound of using the data twice for both item calibration and latent trait estimation. We randomly selected 15,000 individuals as a calibration sample to estimate item parameters for a five factor graded response model. Each item had five item response categories (where 1=Disagree, 3=Neutral, 5=Agree). We used the remainder to apply person fit analysis to detect individuals with potential aberrant behavior. Example code can be found in the supplementary material.

Table 8 presents the proportion of individuals identified as aberrant individuals across each subscale using l_z or l^*_z and the entire scale using l_{zm} or l^*_{zm}. Clearly, using only a subscale resulted in fewer individuals identified as having aberrant behavior where the average proportion of participants identified was 0.079 and 0.114 using l_z and l^*_z compared to 0.125 and 0.165 using l_{zm} and l^*_{zm}. Moreover, applying any asymptotic correction resulted in more individuals flagged. Both of these findings align with the simulation study.

TABLE 9.
Correlation between detected individuals using the Big Five data set.

	Openness	Consciousness	Extraversion	Agreeableness	Neuroticism	Total
Consciousness	1.000	0.266	0.228	0.224	0.215	0.472
Consciousness	0.266	1.000	0.302	0.282	0.277	0.506
Extraversion	0.228	0.302	1.000	0.259	0.276	0.481
Agreeableness	0.224	0.282	0.259	1.000	0.256	0.473
Neuroticism	0.215	0.277	0.276	0.256	1.000	0.518
Total	0.472	0.506	0.481	0.473	0.518	1.000

Table 9 presents the correlation between identified individuals for each subscale and the total scale using l_z^* for each subscale or l_{zm}^*. The average correlation between individual subscales was 0.258, which suggests that there are different sets of individuals being identified by each subscale using l_z^*. However, the correlations between each subscale using l_z^* and the total scale using l_{zm}^* were larger, where the average was 0.490. This suggests that there is more overlap between individuals identified by l_{zm}^* and l_z^* for each subscale. However, there remains certain individuals identified by l_{zm}^* which were not flagged by analyzing only the subscales. This could be because l_{zm}^* is able to flag more aberrant individuals, as demonstrated with the simulations.

Figure 2 shows an example of four individuals who were flagged by l_{zm}^*. Participant 1 was identified by l_{zm}^* and some of the subscale-level l_z^*. Participant 1 appears to have aberrant responses after the 20^{th} item. Prior to the 20^{th} item, the responses appear to be consistent in certain item categories. However, the item responses fluctuate more after the 20^{th} item. This provides corroborating evidence that there may be a fatigue effect after the 20^{th} item (Jeon & De Boeck, (2019)). Participant 2 was identified by all subscale-level l_z^* and l_{zm}^*, where a middle heaping pattern may be occuring. Participant 3 and 4 were identified only by l_{zm}^*, where they exhibited random behavior and extreme response styles across the entire survey.

The applied analysis corroborates with the findings based on the simulation results, where l_{zm}^* appears to be more powerful than l_z^* when detecting aberrant individuals, as evidenced by l_{zm}^* identifying more suspect individuals. Our applied analysis also demonstrates how l_{zm}^* can detect a wide variety of aberrant behavior, such as different response styles, fatigue effects, or random responses.

5. Conclusion and Discussion

The current paper proposes an extension of person fit statistics to accommodate multidimensional scales with mixed item types, l_{zm}^*. Given the increasing attention on data quality, the extension helps applied researchers identify individuals with suspect response behaviors. Snijders (2001) proposed the original correction for dichotomous and unidimensional tests. Sinharay (2016a) proposed a correction for polytomous and unidimensional scales. Albers et al. (2016) proposed corrections for dichotomous items and scales measuring multiple latent traits. The current paper fills the gap for corrections of the more general case, multidimensional scales with mixed item types.

The current study also builds on the recommendations made from previous studies. Albers et al. (2016) did not consider the impact of multiple latent traits on dichotomous items, they fixed the number of latent dimensions to be four in their simulation studies. Our simulations suggest that measuring more constructs poses challenges for the detection of aberrant responses.

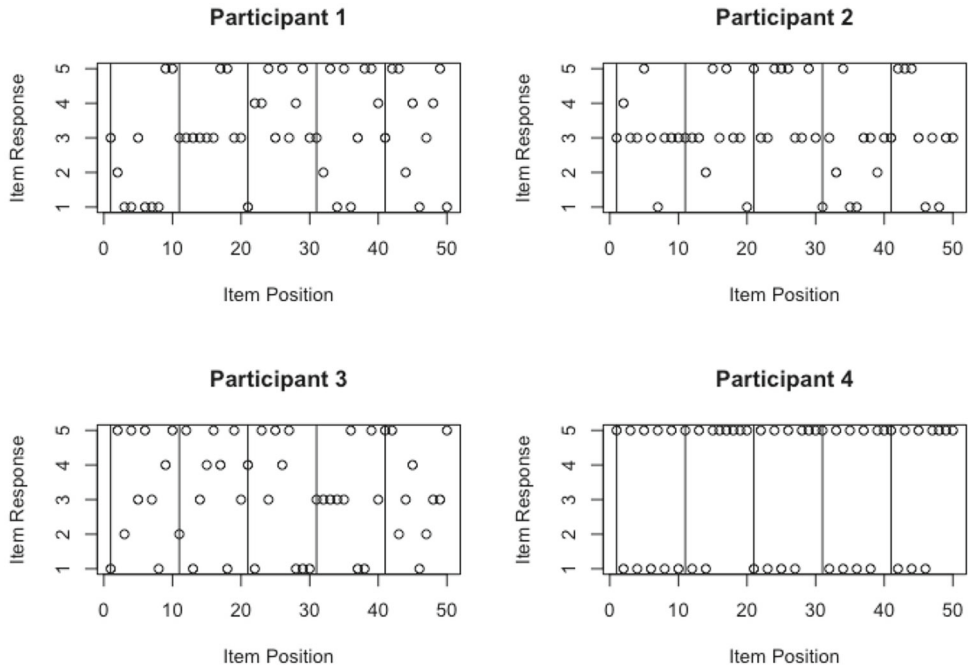


FIGURE 2.
Item response patterns for four aberrant individuals identified by l_{zm}^*. Note: Black lines delineate between subtests.

Moreover, there is a difference when analyzing aberrant responses for individual subscales or an entire survey. When we analyze individual subscales in a survey, we can only detect aberrant behavior inflecting those items. However, aberrant behavior most likely happens at the survey level, such as the fatigue effect. When we analyze aberrant behavior at the survey level, we obtain a more complete picture of each participant. Sinharay (2016a) also did not consider the impact of categories. Tendeiro (2017) found that increased number of categories leads to more conservative type I errors. For example, he found an average type I error rate of 0.03 for l_z^* with polytomous unidimensional scales. Our simulation studies support this finding.

Our applied analysis highlights the utility of analyzing person fit at the survey level in order to detect aberrant individuals. For instance, one type of aberrant behavior that is notorious to detect is when there is a large amount of random responses within a single response vector (Hong et al., 2019). Because we have more information from multiple latent traits, it appears that we can detect these individuals better rather than analyzing subscales individually. Moreover, there appears to be a better ability in detecting nuanced types of aberrant behavior, where analyzing the subtests one at a time misses the complete picture. For instance, the fatigue effect would have not been detected examining the subtests individually.

It is important to note that l_{zm}^* only tests if an individual's response pattern conforms to the assumed measurement model. Individuals need to exhibit aberrant behavior on multiple items for the misfit to be pronounced enough to be caught. If an individual responds carelessly to a handful of items, it is very difficult to detect such a case with person fit. Other statistics may be more appropriate in that case. For instance, change point analysis has become a popular approach to detective changes in response behavior (Shao, Li, & Cheng, 2016; Sinharay, 2016b; Yu & Cheng, 2019). However, these types of statistics, among others (Meijer & Sijtsma, 2001), also require an estimate of a latent trait, which means these alternative approaches would also suffer from the same problems without some correction due to uncertainty introduced by estimated latent trait(s).

There are several limitations to the current paper. For instance, item parameters were assumed to be known in this study. Even if one has good item parameter estimates, uncertainty of the item parameters carries over to the estimation of latent traits, which eventually affects the distribution of the person fit statistics (Cheng & Yuan, 2010). The current derivations can be extended when fallible item parameters are considered. Moreover, the current simulations only considered when the latent trait is estimated with ML. Previous work has found that different estimators did not change the results that much (Sinharay, 2016a); however, more work needs to be done in light of multidimensional constructs. Corrections for two other popular estimators are provided in "Appendix A," and their performances should be evaluated in simulation studies. Furthermore, interpreting the impact of aberrant behavior could be confounded when we consider varying categories. The development of a within person effect size measure for aberrant behavior is a possible future direction. Finally, the current derivations only consider when there is simple structure in the scale configuration. Person fit statistics can be extended to multidimensional constructs with complex structure. Complex structure means there may be more than one underlying latent trait per item. An interesting future direction would be to generalize the current corrections to more general scenarios such as complex structure.

6. Appendix A

6.1. Formulas for Different Estimators of $\boldsymbol{\theta}$: $\boldsymbol{\theta}_{MAP}$ and $\boldsymbol{\theta}_{WLE}$

The following section is based on work done by Sinharay (2016a) and Wang (2015). Suppose $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ satisfies the following condition:

$$\nabla l(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}} = \begin{bmatrix} t_{01}(\boldsymbol{\theta}) \\ t_{02}(\boldsymbol{\theta}) \\ \vdots \\ t_{0S}(\boldsymbol{\theta}) \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij1}(\boldsymbol{\theta}) \\ \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij2}(\boldsymbol{\theta}) \\ \vdots \\ \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ijS}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{63}$$

which can be rewritten as:

$$\nabla l(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}} = t_0(\boldsymbol{\theta}) + \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)t_{ij}(\boldsymbol{\theta}) = \mathbf{0} \tag{64}$$

for some functions $t_0(\boldsymbol{\theta}) = (t_{01}, t_{02}, \ldots, t_{0S})'$ and $t_{ij}(\boldsymbol{\theta}) = (t_{ij1}, t_{ij2}, \ldots, t_{ijS})'$. For instance, $\hat{\boldsymbol{\theta}}_{ML}$ is the value of $\boldsymbol{\theta}$ for which:

$$\nabla l(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}} = \sum_{i=1}^{p}\sum_{j=0}^{m_i}\Big(\mathbb{I}_j(X_i) - P_{ij}(\boldsymbol{\theta})\Big)P_{ij}^{-1}(\boldsymbol{\theta})\nabla P_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}} = \mathbf{0}. \tag{65}$$

The equality in Equation (64) holds where $t_0(\boldsymbol{\theta})$ and $t_{ij}(\boldsymbol{\theta})$ satisfy:

$$t_0(\boldsymbol{\theta}) = \mathbf{0} \text{ and } t_{ij}(\boldsymbol{\theta}) = P_{ij}^{-1}(\boldsymbol{\theta})\nabla P_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}}. \tag{66}$$

Similarly, $\hat{\boldsymbol{\theta}}_{MAP}$ satisfies the following:

$$\nabla l(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}} + \nabla \log \pi(\boldsymbol{\theta})|_{\boldsymbol{\theta}} = \mathbf{0}, \tag{67}$$

where $\pi(\boldsymbol{\theta})$ is a prior distribution for $\boldsymbol{\theta}$. Equation (64) holds for $\hat{\boldsymbol{\theta}}_{MAP}$ where:

$$t_0(\boldsymbol{\theta}) = \nabla \log \pi(\boldsymbol{\theta})|_{\boldsymbol{\theta}} \text{ and } t_{ij}(\boldsymbol{\theta}) = P_{ij}^{-1}(\boldsymbol{\theta}) \nabla P_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}}. \tag{68}$$

Note that if the prior is a standard multivariate normal distribution, then $\nabla \log \pi(\boldsymbol{\theta})|_{\boldsymbol{\theta}} = -\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}_{WLE}$ satisfies Equation (64) where:

$$\nabla l(X|\boldsymbol{\theta})|_{\boldsymbol{\theta}} + \nabla \bar{\mathbf{I}}_p(\boldsymbol{\theta}) \mathbf{B}(\boldsymbol{\theta})|_{\boldsymbol{\theta}} = \mathbf{0}, \tag{69}$$

where $\bar{\mathbf{I}}_p$ be the average information about $\boldsymbol{\theta}$ in the sample where $\bar{\mathbf{I}}_p = \sum_{i=1}^{p} \mathbf{I}_i(\boldsymbol{\theta})/p$. $\mathbf{B}(\boldsymbol{\theta}) = [B(\theta_1), B(\theta_2), ..., B(\theta_S)]'$ is a S-dimensional vector where the s^{th} element in $\mathbf{B}(\boldsymbol{\theta})$ is:

$$\frac{1}{2} \sum_{t,u,v=1}^{p} I^{uv} I^{vt} E\Big(\frac{\partial^3 l}{\partial \theta_t \partial \theta_u \partial \theta_v}\Big) \tag{70}$$

Therefore this satisfies Equation (64) where

$$t_0(\boldsymbol{\theta}) = \nabla \bar{\mathbf{I}}_p(\boldsymbol{\theta}) \mathbf{B}(\boldsymbol{\theta})|_{\boldsymbol{\theta}} \text{ and } t_{ij}(\boldsymbol{\theta}) = P_{ij}^{-1}(\boldsymbol{\theta}) \nabla P_{ij}(\boldsymbol{\theta})|_{\boldsymbol{\theta}}. \tag{71}$$

References

Albers, C. J., Meijer, R. R., & Tendeiro, J. N. (2016). Derivation and applicability of asymptotic results for multiple subtests person-fit statistics. *Applied Psychological Measurement, 40*(4), 274–288. https://doi.org/10.1177/0146621615622832.

Baer, R. A., Ballenger, J., Berru, D., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*(1), 139–151.

Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the rasch model. *Psychometrika, 62,* 191–199. https://doi.org/10.1007/BF02295274.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*(3), 340–345. https://doi.org/10.1037/1040-3590.4.3.340.

Bhattacharya, R., Lin, L., & Victor, P. (2016). *A course in mathematical statistics and large sample theory*. Berlin: Springer.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F. M. Lord & M. Novick (Eds.), Statistical theories of mental test scores (pp 397–472).

Casella, G. & Berger, R. (2001). Statistical Inference (No. 141). https://doi.org/10.1057/pt.2010.23

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software 48 (6). Retrieved from http://www.jstatsoft.org/v48/i06/ https://doi.org/10.18637/jss.v048.i06

Cheng, Y., & Yuan, K. H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75,* 280–291. https://doi.org/10.1007/s11336-009-9144-x.

Conijn, J. M., Emons, W. H., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*(2), 122–136. https://doi.org/10.1177/0146621613497568.

Conrad, K. J., Bezruczko, N., Chan, Y. F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence, 106,* 92–100. https://doi.org/10.1016/j.drugalcdep.2009.07.023.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*(2), 171–191. https://doi.org/10.1177/014662169101500207.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26.
Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions. Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*(1), 82–98. https://doi.org/10.1037/0022-3514.48.1.82.
Hong, M., Steedle, J. T., & Cheng, Y. (2019). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80*(2), 312–345. https://doi.org/10.1177/0013164419865316.
Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology, 72*(3), 517–537. https://doi.org/10.1111/bmsp.12182.
Karabatsos, G. (2003). *Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics., 16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604.
Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of snijders's l z* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*(1), 57–81. https://doi.org/10.3102/1076998610396894.
Magnus, J., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.
Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*(1), 1–2.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135. (**Retrieved from**).
Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75–106. https://doi.org/10.1007/BF02294745.
Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63,* 1–11. https://doi.org/10.1016/j.jrp.2016.04.010.
Reckase, M. (2009). *Multidimensional Item Response Theory.* https://doi.org/10.1007/978-0-387-89976-3.
Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127–137. https://doi.org/10.1177/014662169001400202.
Rizopoulos, D. (2006). Itm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software.* https://doi.org/10.18637/jss.v017.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*(1), 3–38.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Vol. 35) (No. 1). https://doi.org/10.1007/BF02290599
Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81*(4), 1118–1141. https://doi.org/10.1007/s11336-015-9476-7.
Sinharay, S. (2016a). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81,* 992–1013. https://doi.org/10.1007/s11336-015-9465-x.
Sinharay, S. (2016b). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika, 82,* 1–13. https://doi.org/10.1007/s11336-016-9531-z.
Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342. https://doi.org/10.1007/BF02294437.
Tendeiro, J. N. (2017). The lz(p)* person-fit statistic in an unfolding model context. *Applied Psychological Measurement, 41*(1), 44–59. https://doi.org/10.1177/0146621616669336.
Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). Perfit: An R package for person-fit analysis in IRT. *Journal of Statistical Software, 74*(5), 1–27. https://doi.org/10.18637/jss.v074.i05.
von Davier, M., & Molenaar, I. W. (2003). A person-fit index for polytomous rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68,* 213–228. https://doi.org/10.1007/BF02294798.
Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika, 80,* 428–449. https://doi.org/10.1007/s11336-013-9399-0.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450. https://doi.org/10.1007/BF02294627.
Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 5,* 658–674. https://doi.org/10.1037/met0000212.
Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249. https://doi.org/10.1007/BF02294536.

Manuscript Received: 4 MAY 2020
Final Version Received: 3 NOV 2020
Accepted: 26 FEB 2021
Published Online Date: 1 APR 2021