PSYCHOMETRIKA—VOL. 89, NO. 4, 1230–1260
DECEMBER 2024
https://doi.org/10.1007/s11336-024-09997-y

ASYMPTOTICALLY CORRECT PERSON FIT Z-STATISTICS FOR THE RASCH TESTLET MODEL

ZHONGTIAN LIN

FINANCIAL INDUSTRY REGULATORY AUTHORITY

TAO JIANG, FRANK RIJMEN AND PAUL VAN WAMELEN

CAMBIUM ASSESSMENT

A well-known person fit statistic in the item response theory (IRT) literature is the l_z statistic (Drasgow et al. in Br J Math Stat Psychol 38(1):67-86, 1985). Snijders (Psychometrika 66(3):331-342, 2001) derived l_z^*, which is the asymptotically correct version of l_z when the ability parameter is estimated. However, both statistics and other extensions later developed concern either only the unidimensional IRT models or multidimensional models that require a joint estimate of latent traits across all the dimensions. Considering a marginalized maximum likelihood ability estimator, this paper proposes l_{zt} and l_{zt}^*, which are extensions of l_z and l_z^*, respectively, for the Rasch testlet model. The computation of l_{zt}^* relies on several extensions of the Lord-Wingersky algorithm (1984) that are additional contributions of this paper. Simulation results show that l_{zt}^* has close-to-nominal Type I error rates and satisfactory power for detecting aberrant responses. For unidimensional models, l_{zt} and l_{zt}^* reduce to l_z and l_z^*, respectively, and therefore allows for the evaluation of person fit with a wider range of IRT models. A real data application is presented to show the utility of the proposed statistics for a test with an underlying structure that consists of both the traditional unidimensional component and the Rasch testlet component.

Key words: Person fit, IRT, l_z statistic, Rasch testlet model.

Item response theory (IRT) is ubiquitously used as the underlying statistical model for calibrating items and scoring examinee responses. Establishing that the IRT model adequately fits the data is an important aspect of establishing validity for the intended use of test scores resulting from the assessment. Item statistics, including item fit, are taken into account during the item review process and could indicate that the item should be modified or rejected altogether. Even when all items fit the model, it is possible that the model does not fit for a particular examinee. For example, a person answering all easy items incorrectly but all other items correctly is an unexpected or aberrant response pattern for a given set of item parameters. Aberrant responses refer to a series of answers examinees provided that are unlikely to arise given their true ability and the chosen psychometric model. In other words, there is a lack of fit between response patterns and the model used for scoring. Many test-taking behaviors such as cheating, lack of motivation and random response can cause aberrant responses and lead to a poor person fit.

Various indices have been proposed to capture the degree of person fit (see Meijer and Sijtsma, 2001, or Karabatsos, 2003, for a survey of person fit statistics in the earlier literature. More recently, fit statistics were proposed by, among others, von Davier and Molenaar 2003; Glas and Dagohoy 2007; de la Torre and Deng and Deng, 2008; Sinharay 2015; 2016; Xia and Zheng, 2018. A relatively recent review can be found in Rupp, 2013). Among these statistics, one of the most

The research reported in this paper was performed when the first author was an employee of Cambium Assessment. The first author is currently an employee of the Financial Industry Regulatory Authority (FINRA). Any opinion expressed in this publication are those of the authors and not necessarily of FINRA

Correspondence should be made to Zhongtian Lin, Financial Industry Regulatory Authority, Washington, USA. Email: lzt713@gmail.com

well-known is the standardized loglikelihood statistic of a response pattern, denoted as l_z, first developed by (Drasgow et al., 1985) . The statistic provides a measure of the degree to which the response pattern is aberrant, given a known value of the true ability (θ). l_z asymptotically follows a standard normal distribution (Drasgow et al., 1985; Snijders, 2001) .

In practice however, the true ability is not known but estimated from the same data on which l_z is computed. Even though Drasgow et al. (1985) indicated that the effects from the estimated ability ($\hat{\theta}$) were fairly small given the fact that standardization of the response loglikelihood has reduced its dependency on the estimated ability, other researchers have found scenarios where l_z deviates from standard normal. Molenaar and Hoijtink (1990) found that in Rasch model for dichotomous items, even when assuming $\hat{\theta} = \theta$ given a raw score, the deviation from normality of l_z was particularly evident when $\hat{\theta}$ was far from the mean of the item difficulties, and when the test was short. Negative skewness and heavy tails were observed in the example cases they showed. Several other studies have found that the variance of l_z can be considerably smaller than 1 when the true ability θ is replaced by the estimated ability ($\hat{\theta}$) (e.g., Nering, 1995; Reise, 1995; Seo & Weiss, 2013). Molenaar and Hoijtink (1990) proposed a modified version of the person fit index, by using the result that the sum of the raw scores is a sufficient statistic for $\hat{\theta}$ for the Rasch model. The first few central moments of the proposed statistic were computed and used in deriving a chi-squared distribution-based approximation that accounts for the skewness of the loglikelihood person fit index. Bedrick (1997) used a different approximation that involves the use of Edgeworth expansion for skewness correction. von Davier and Molenaar (2003) extended the work of Molenaar and Hoijtink (1990) to latent class models and mixture distribution IRT models for both dichotomous and polytomous data. They also compared the performance of the two aforementioned approaches to reduce the skewness of the person-fit index. Liou and Chang (1992), on the other hand, used a so-called network algorithm to obtain the exact significance of the loglikelihood person fit index when conditioning on either the maximum likelihood ability estimates or the sum score in Rasch model. Meanwhile, Snijders (2001) derived a framework of asymptotically normal person fit statistics for dichotomous items when the $\hat{\theta}$ is used, among which is the modified version of l_z now commonly referred to as the l_z^* statistic. When $\hat{\theta}$ is the maximum likelihood estimate, the essence of l_z^* lies in correcting the loglikelihood variance estimate in the original of l_z. It was shown in Snijders (2001) that l_z^* produced type I error rates close to the nominal rate. Sinharay (2016) derived l_z^* for mixed format test, where polytomous items can also be handled along with dichotomous items.

An important limitation of l_z and l_z^*, and the other previously mentioned person fit indices in the literature, is that they only address the person fit assessment with a unidimensional latent trait. More recently, there have been some efforts to extend l_z and l_z^* for their uses with multidimensional constructs. Albers et al. (2016) proposed l_{zm} and l_{zm}^*, which are used for dichotomous items and multiple subscales. Hong et al. (2121) provided more rigorous derivations of these statistics, extensions to mixed item types, and more extensive simulation studies. It should be noted that an implicit requirement to use l_{zm} or l_{zm}^* is that person estimates are obtained across all dimensions. In practice, however, one of the important use cases of introducing additional latent variables is to address the local dependencies among items that share a common stimulus or belong to the same testlet. Some popular models developed to this end are the testlet models (Bradlow et al., 1999) and particularly the widely used Rasch testlet model (Wang & Wilson, 2005) . In such cases, usually the overarching latent trait is of primary interest, while the other traits are incorporated as so-called "nuisance" dimensions to account for the testlet effects. When examining the person fit with these models, l_{zm} or l_{zm}^* cannot be applied unless θ estimates for all the dimensions are obtained, counter to the idea of introducing testlet effects as nuisance dimensions. On the other hand, a direct application of l_z and l_z^* ignoring the testlet effects is also not a good solution. Chen (2013) investigated the utility of l_z on detecting aberrant responses for the testlet model and found that the detection rate was worse when there were more testlet items or the testlet variance was

larger. In sum, there is a need to develop a feasible approach of person fit evaluation that works for testlet models.

This paper proposes two new statistics, l_{zt} and l_{zt}^*, which extend l_z and l_z^*, respectively for the Rasch testlet model when the marginalized maximum likelihood estimation (MMLE) is used for θ estimation (i.e., the nuisance dimensions are integrated out. More details about MMLE are provided in a later section of this paper). Moreover, with the advances in technology enhanced items and test delivery system, test developers nowadays create novel tests with an underlying latent structure that incorporates both items organized in testlets and unidimensional standalone items (e.g., New Hampshire Department of Education, 2019). It will be shown that l_{zt} and l_{zt}^* reduce to l_z and l_z^*, respectively, with unidimensional MLE estimation of θ, and therefore can be considered as a generalized approach to evaluate person fit when the underlying structure includes both a testlet component and observed variables that do not belong to any testlet.

The rest of this paper is organized in the following way. First, we provide some theoretically background on l_z and l_z^*, as well as some technical details about the MMLE method for the estimation of the overall θ under the Rasch testlet model. We then extend the original l_z statistics to its form in the Rasch testlet model and illustrate how the variance of the loglikelihood can be corrected when MMLE is used to obtain the new statistics we call l_{zt}^*. A simulation study follows to evaluate the performance of l_{zt}^*, including the Type I error rate and power under the Rasch testlet model. We then demonstrate an application of l_{zt}^* on a real dataset from a large-scale standardized assessment, to show that l_{zt}^* is flexible such that it can be applied to a wider range of models which allow for both the testlet model for some item sets and a traditional unidimensional model for other items. Finally, we discuss practical considerations and future direction of these statistics.

1. Review of the l_z and l_z^* Statistics for Unidimensional Models

Because the extension of l_z and l_z^* this paper presents mainly concerns the Rasch testlet model for dichotomous item responses, we offer a review of l_z and l_z^* for dichotomous items here to achieve a better connection to the method to be proposed. A didactic presentation of l_z^* was offered by Magis, Raîche, and Béland (2012). A presentation of l_z and l_z^* for mixed format tests is available from Sinharay (2016) where l_z^* for dichotomous items was shown as a special case.

Consider an examinee with true ability θ who responds to a test consists of n items modeled by a unidimensional IRT model (for example, the one-, two-, and three-parameter logistic model). Throughout the paper, item parameters of the IRT models are assumed to be known. Let Y_j be the binary response provided by the examinee to item j, $p_j(\theta) = P(Y_j = 1|\theta)$ be the probability of correct response to item j, and $q_j(\theta) = 1 - p_j(\theta)$. As defined by Snijders (2001), one class of the person fit statistics W_j for dichotomous items can be expressed in a centered form as

$$W(\theta) = \sum_{j=1}^{n} \left(Y_j - p_j(\theta)\right) w_j(\theta),$$

where $w_j(\theta)$ is a suitable weight function. The random variance $W(\theta)$ has expected value

$$E(W(\theta)) = 0$$

and variance

$$Var(W(\theta)) = n\sigma_n^2(\theta) = \sum_{j}^{n} w_j(\theta)\, p_j(\theta)\, q_j(\theta).$$

Under regularity conditions, the standardized version of $W(\theta)$ which takes the form

$$\frac{W(\theta)}{Var(W(\theta))}$$

asymptotically follows a standard normal distribution by the Lindeberg-Feller central limit theorem for independent but non-identically distributed random variables. The l_z statistics (Drasgow et al., 1985) is defined as

$$l_z(\theta) = \frac{l(\theta) - E(l(\theta))}{Var(l(\theta))}. \tag{1}$$

For dichotomous items,

$$l(\theta) = \sum_j^n Y_j \log p_j(\theta) + (1 - Y_j) \log q_j(\theta),$$

which is the log-likelihood of the examinee's item scores. The expected value of $l(\theta)$ is

$$E(l(\theta)) = \sum_j^n p_j(\theta) \log p_j(\theta) + q_j(\theta) \log q_j(\theta),$$

and the variance of $l(\theta)$ is

$$Var(l(\theta)) = p_j(\theta) q_j(\theta) \left(\log \frac{p_j(\theta)}{q_j(\theta)}\right)^2.$$

$l_z(\theta)$ is a special case of the standardized version of $W(\theta)$ when

$$w_j(\theta) = \log \frac{p_j(\theta)}{q_j(\theta)}.$$

Note that $W(\theta)$ (or $l_z(\theta)$) is defined in terms of true ability θ. However, when applied to real data, θ is unknown and must be replaced by the estimated value $\hat{\theta}$. Several research studies have shown that $l_z\left(\hat{\theta}\right)$ differs from a standard normal distribution when $\hat{\theta}$ is used and therefore provides an inaccurate assessment of person fit (Molenaar & Hoijtink 1990; Nering, 1995; Reise 1995; Snijders, 2001; van Krimpen-Stoop & Meijer, 1999). Snijders (2001) provided a remedy to this problem. First, using the Taylor expansion on $W(\theta)$, he showed

$$\frac{1}{\sqrt{n}} W\left(\hat{\theta}\right) \approx \frac{1}{\sqrt{n}} W(\theta) + \sqrt{n}\left(\hat{\theta} - \theta\right) \left[\frac{1}{n} \sum_{j=1}^n \left(Y_j - p_j(\theta)\right) w_j'(\theta) - \frac{1}{n} \sum_{j=1}^n p_j'(\theta) w_j(\theta)\right],$$

where $w_j'(\theta)$ and $p_j'(\theta)$ are the first derivative of $w_j(\theta)$ and $p_j(\theta)$, respectively. The term $\sqrt{n}\left(\hat{\theta} - \theta\right)$ is bounded assuming it has a non-degenerate distribution when $n \to \infty$. While the first term in the bracket tends to 0 since it is an average of a random variable with expected value

of 0, the second term in the bracket does not. Snijders suggested to replace $w_j(\theta)$ with a $\tilde{w}_j(\theta)$ such that $\sum_{j=1}^{n} p_j^{'}(\theta)\,\tilde{w}_j(\theta) = 0$. To be specific, if a $\hat{\theta}$ satisfies the condition that

$$r_0\left(\hat{\theta}\right) + \sum_{j=1}^{n}\left(Y_j - p_j\left(\hat{\theta}\right)\right) r_j\left(\hat{\theta}\right) = 0.$$

The modified weight $\tilde{w}_j(\theta)$ can be defined as

$$\tilde{w}_j(\theta) = w_j(\theta) - c_n(\theta)\, r_j(\theta), \tag{2}$$

where

$$c_n(\theta) = \frac{\sum_{j=1}^{n} p_j^{'}(\theta)\, w_j(\theta)}{\sum_{j=1}^{n} p_j^{'}(\theta)\, r_j(\theta)}.$$

Then, the new variable

$$W^*\left(\hat{\theta}\right) = \frac{W\left(\hat{\theta}\right) + c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)}{Var\left(W^*\left(\hat{\theta}\right)\right)}$$

asymptotically follows a standard normal distribution, where

$$Var\left(W^*\left(\hat{\theta}\right)\right) = n\tau_n\left(\hat{\theta}\right) = \sum_{j=1}^{n} \tilde{w}_j^2\left(\hat{\theta}\right) p_j\left(\hat{\theta}\right) q_j\left(\hat{\theta}\right) \quad . \tag{3}$$

For an MLE, $r_0\left(\hat{\theta}\right) = 0$; For a maximum a posteriori (MAP) estimator, $r_0\left(\hat{\theta}\right) = d\log\left(f\left(\hat{\theta}\right)\right)/d(\hat{\theta})$, where $f\left(\hat{\theta}\right)$ is a prior distribution of ability; For a weighted likelihood estimator (WLE), $r_0\left(\hat{\theta}\right) = J\left(\hat{\theta}\right)/2\left(I\left(\hat{\theta}\right)\right)$, where $J\left(\hat{\theta}\right) = \sum_{j=1}^{n} \frac{p_j^{'}\left(\hat{\theta}\right) p_j^{''}\left(\hat{\theta}\right)}{p_j\left(\hat{\theta}\right) q_j\left(\hat{\theta}\right)}$, $I\left(\hat{\theta}\right) = \sum_{j=1}^{n} \frac{p_j^{'}\left(\hat{\theta}\right)^2}{p_j\left(\hat{\theta}\right) q_j\left(\hat{\theta}\right)}$ and $p_j^{''}(\theta)$ is the second derivative of $p_j(\theta)$. $r_j\left(\hat{\theta}\right)$ is given in general by

$$r_j\left(\hat{\theta}\right) = \frac{p_j^{'}\left(\hat{\theta}\right)}{p_j\left(\hat{\theta}\right) q_j\left(\hat{\theta}\right)}$$

Consequently,

$$l_z^*\left(\hat{\theta}\right) = \frac{l\left(\hat{\theta}\right) - E\left(l\left(\hat{\theta}\right)\right) + c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)}{Var\left(l_z^*\left(\hat{\theta}\right)\right)}. \tag{4}$$

Comparing Eq. (1) with Eq. (4), we see that $l_z^*\left(\hat{\theta}\right)$ is obtained using the equation of $l_z\left(\hat{\theta}\right)$ by adjusting the mean with $c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)$ and adjusting the variance by replacing $Var\left(l_z\left(\hat{\theta}\right)\right)$ with $Var\left(l_z^*\left(\hat{\theta}\right)\right)$. Particularly for an MLE, since $r_0\left(\hat{\theta}\right) = 0$, only the variance needs to be adjusted, and the above formula reduces to

$$l_z^*\left(\hat{\theta}\right) = \frac{l\left(\hat{\theta}\right) - E\left(l\left(\hat{\theta}\right)\right)}{Var\left(l_z^*\left(\hat{\theta}\right)\right)} \quad . \tag{5}$$

As we will show later in the *Method* section, this adjustment of the variance under MLE is a general strategy on which we relied when adjusting the extended version of l_z for the Rasch testlet model under MMLE. To provide a better connection, we shall now take a closer look at $Var\left(l_z^*\left(\hat{\theta}\right)\right)$ to see what information is needed to compute it. Omitting $\hat{\theta}$ for simplicity, based on Eqs. (2) and (3), we have

$$Var\left(l_z^*\right) = \sum_{j=1}^{n}\left(w_j - c_n r_j\right)^2 p_j q_j,$$

where $c_n = \frac{\sum_{j=1}^{n} p_j^{'} w_j}{\sum_{j=1}^{n} p_j^{'} r_j}$, $r_j = \frac{p_j^{'}}{p_j q_j}$ and $w_j = \log\frac{p_j}{q_j}$. Therefore

$$\begin{aligned} Var\left(l_z^*\right) &= \sum_{j=1}^{n}\left(\log\frac{p_j}{q_j} - \left(\frac{\sum_{j=1}^{n} p_j^{'}\log\frac{p_j}{q_j}}{\sum_{j=1}^{n}\frac{p_j^{'2}}{p_j q_j}}\right)\frac{p_j^{'}}{p_j q_j}\right)^2 p_j q_j \\ &= \sum_{j=1}^{n} p_j q_j\left(\log\frac{p_j}{q_j}\right)^2 - 2\left(\sum_{j=1}^{n} p_j^{'}\log\frac{p_j}{q_j}\right) * \frac{\sum_{j=1}^{n} p_j^{'}\log\frac{p_j}{q_j}}{\sum_{j=1}^{n}\frac{p_j^{'2}}{p_j q_j}} + \frac{\left(\sum_{j=1}^{n} p_j^{'}\log\frac{p_j}{q_j}\right)^2}{\sum_{j=1}^{n}\frac{p_j^{'2}}{p_j q_j}} \\ &= \sum_{j=1}^{n} p_j q_j\left(\log\frac{p_j}{q_j}\right)^2 - \frac{\left(\sum_{j=1}^{n} p_j^{'}\log\frac{p_j}{q_j}\right)^2}{\sum_{j=1}^{n}\frac{p_j^{'2}}{p_j q_j}}. \end{aligned}$$

We should now examine the terms of the final form of $Var\left(l_z^*\right)$ above. The first term is exactly the original definition of $Var\left(l_z\right)$. For the numerator of the second term, if we define $h\left(\hat{\theta}\right) = l\left(\hat{\theta}\right) - E\left(l\left(\hat{\theta}\right)\right)$ (note that this is the numerator of l_z^*), we find it amounts to $\left(h^{'}\left(\hat{\theta}\right)\right)^2$ for an MLE $\hat{\theta}$, where $h^{'}\left(\hat{\theta}\right) = -\sum_{j=1}^{n}\left(p_j^{'}\log\frac{p_j}{q_j}\right)$ is the first derivative of $h\left(\hat{\theta}\right)$. Finally, the denominator of the second term can be recognized as test information at $\theta = \hat{\theta}$ (let's denote it as $I\left(\hat{\theta}\right)$). Therefore, we can rewrite the above definition of $Var\left(l_z^*\right)$ as

$$Var\left(l_z^*\left(\hat{\theta}\right)\right) = Var\left(l_z\left(\hat{\theta}\right)\right) - \frac{\left(h^{'}\left(\hat{\theta}\right)\right)^2}{I\left(\hat{\theta}\right)}.$$

This alternative definition of $Var\left(l_z^*\right)$, as we shall see in the later section of this paper, holds true when l_z^* is extended for the Rasch testlet model.

2. Rasch Testlet Model and MMLE θ Estimation

Before we describe our extended method, we provide some basic information about the Rasch testlet model and the utility of MMLE estimation of θ. While unidimensional models have been working well with tests that consist of traditional items, it is arguably not the best choice when a test consists of testlets. A testlet, sometime called an item cluster or an item bundle, is a set of items that share a common stimulus. Because of such bundling, an examinee's responses to items within a testlet are usually interdependent even when conditioned on the examinee ability. That is, the usual local independence assumption does not hold within testlets. Ignoring such dependencies would result in biased item parameter estimates and underestimation of the standard error of measurement (e.g., Sireci et al., 1991; Wainer & Lukhele, 1997; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993). A common approach to account for the testlet effect is to include additional dimensions corresponding to the bundling of the items in the IRT model. These additional dimensions incorporated are usually considered "nuisance" dimensions as the true values of examinees' latent traits on these dimensions are often not of primary interest. One popular example of adopting this approach is the Rasch testlet model. For binary data, the Rasch testlet model is defined as

$$p_{jk}\left(\theta|u_k\right) = P\left(Y_{jk}=1|\theta,\, u_k\right) = \frac{\exp\left(\theta+u_k-b_j\right)}{1+\exp\left(\theta+u_k-b_j\right)}, \tag{6}$$

where Y_{jk} is the response to item j from testlet k and can be either 0 or 1, θ is the examinee's overall ability, u_k is the latent trait related to testlet k, and b_j is the difficulty parameter of item j.

To understand how l_z and l_z^* can be extended for the Rasch testlet model, there is a need to review the methods for the estimation of latent traits in multidimensional IRT (MIRT) models. Two commonly used estimators for the latent traits in MIRT models are the maximum likelihood estimator (MLE) and the expected a posteriori (EAP) estimator. Let y be a vector collecting the observed item scores for all items in all testlets, and u be a vector collecting the latent traits pertain to the nuisance dimensions. The MLE is obtained by maximizing the likelihood of the observed items scores jointly for θ and u. That is,

$$\left(\hat{\theta},\, \hat{\boldsymbol{u}}\right)_{MLE} = \text{argmax}_{\theta,\boldsymbol{u}} l\left(\theta, \boldsymbol{u}|\boldsymbol{y}\right),$$

where $l\left(\theta, \boldsymbol{u}|\boldsymbol{y}\right)$ is the log-likelihood of the observed item scores. The EAP estimator is the posterior mean vector of the latent traits, defined as

$$\left(\hat{\theta},\, \hat{u}\right)_{EAP} = \int_{-\infty}^{\infty} \left(\theta, \boldsymbol{u}\right) p\left(\theta, \boldsymbol{u}|\boldsymbol{y}\right) d\left(\theta, \boldsymbol{u}\right),$$

where $p\left(\theta, \boldsymbol{u}|\boldsymbol{y}\right)$ is the joint posterior distribution of θ and $\boldsymbol{u}$, given the observed item score vector. Both estimators are multivariate, i.e., they jointly obtain the estimate of the overall ability θ and the estimates of the latent traits regarding the testlet effects ($\boldsymbol{u}$). Therefore, when these two methods are used, the log-likelihood involved in obtaining l_z and the corresponding correction involved to obtain l_z^* can be considerably more difficult to disentangle than those in a unidimensional model.

However, the purpose of introducing the nuisance dimensions $\boldsymbol{u}$ is solely to account for the item clustering or testlet effect; Most of the time, only the overall θ is of primary interest. In this vein, Rijmen et al. (2018) proposed to use the marginalized maximum likelihood estimator (MMLE) for the overall θ estimation. The MMLE can be obtained in two steps. First, the nuisance dimensions $\boldsymbol{u}$ are integrated out in the observed data likelihood to obtain the marginalized likelihood function of θ,

$$L\left(\theta|\boldsymbol{y}\right) = \int_{-\infty}^{\infty} p\left(\boldsymbol{y}|\theta, \boldsymbol{u}\right) p\left(\boldsymbol{u}\right) d\boldsymbol{u}.$$

Second, $\hat{\theta}$ is found by maximizing the resulting marginal (log-)likelihood function,

$$\left(\hat{\theta}\right)_{MMLE} = \text{argmax}_{\theta} l_{marginal}\left(\theta|\boldsymbol{y}\right),$$

where $l_{marginal} = Log\left(L\left(\theta|\boldsymbol{y}\right)\right)$. In a simulation study, Rijmen et al. (2018) found that the MMLE provided a better recovery of the overall ability parameter than the MLE and EAP estimators in the presence of substantial testlet effects, and that only the MMLE accurately took into account the loss of information due the dependencies between items from the same stimulus. The mathematical simplicity of MMLE relative to the other joint estimators offers an opportunity to develop a suitable person fit measure on the basis of l_z. The next section shows that the original l_z statistics can be extended to work for the Rasch testlet model, and an asymptotically corrected version can be derived to produce a new person fit z-statistic when the MMLE $\hat{\theta}$ is used.

3. Method

3.1. Extension of l_z for the Rasch testlet model

Consider a test that consists of K testlets where each item within a testlet is scored either 0 or 1. The probability of getting a score of y_{jk} for item j in testlet k based on the Rasch testlet model is defined as

$$p_{y_{jk}} = \text{P}\left(Y_{jk} = y_{jk}|\theta, u_k\right) = \left(p_{jk}\left(\theta|u_k\right)\right)^{y_{jk}} \left(q_{jk}\left(\theta|u_k\right)\right)^{1-y_{jk}},$$

where $p_{jk}\left(\theta|u_k\right)$ is as defined in Eq. (6), and $q_{jk}\left(\theta|u_k\right) = 1 - p_{jk}\left(\theta|u_k\right)$.

The likelihood of the overall ability θ for an MMLE is defined as

$$L\left(\theta|\boldsymbol{y}\right) = \prod_{k=1}^{K} \int \prod_{j=1_k}^{n} \text{P}\left(Y_{jk} = y_{jk}|\theta, u_k\right) g\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k,$$

where n_k is the number of items in testlet k, $g\left(u_k \mid 0,\ \sigma_{u_k}^2\right)$ is the assumed prior distribution of u_k with a mean of 0 and a variance of $\sigma_{u_k}^2$. The log-likelihood statistics is therefore

$$l\left(\theta|\mathbf{y}\right)=\sum_{k=1}^{K}\log\left[\int\text{Exp}\left(\sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}\left(\theta|u_k\right)\right)\right.\right.\right.$$
$$\left.\left.\left.+\left(1-y_{jk}\right)\log\left(q_{jk}\left(\theta|u_k\right)\right)\right)\right)g\left(u_k|0,\ \sigma_{u_k}^2\right)du_k\right].$$

Analogous to what it is in a unidimensional model, the standardized log-likelihood statistics is defined as

$$l_{zt}\left(\theta\right)=\frac{l\left(\theta|\mathbf{y}\right)-E\left(l\left(\theta|\mathbf{y}\right)\right)}{\sqrt{Var\left(l\left(\theta|\mathbf{y}\right)\right)}}.$$

Under regularity conditions, $l_{zt}\left(\theta\right)$ follows a standard normal distribution and can be used for person fit evaluations. The obstacle here is to compute $E\left(l\left(\theta|\mathbf{y}\right)\right)$ and $Var\left(l\left(\theta|\mathbf{y}\right)\right)$. As a model from the Rasch family, a merit of the Rasch testlet model is that a sufficient statistic for θ exists in a relatively simple form. Similar to the unidimensional Rasch model where the sum of the items score of the entire test is a sufficient statistic for θ, Appendix A shows that the vector $\{r_1, r_2, \cdots, r_k,\ r_{k+1}, \cdots, r_K\}$ is the sufficient statistic for θ, where r_k is the sum of the item scores of testlet k and K is the total number of the testlets. Therefore,

$$E\left(l\left(\theta|\mathbf{y}\right)\right)=\sum_{k=1}^{K}\left[E\left(l\left(\theta\,|\,r_k\right)\right)\right].$$

In the equation above,

$$E\left(l\left(\theta|r_k\right)\right)=\int\left(\sum_{j=1}^{n_k}p_{jk}\left(\theta|u_k\right)\left(\theta-b_{jk}\right)\right)g\left(u_k\,|\,0,\ \sigma_{u_k}^2\right)du_k$$
$$+\sum\nolimits_{r_k=0}^{n_k}\left\{\log\left[\int\text{Exp}\left(r_ku_k+\sum_{j=1}^{n_k}\log\left(q_{jk}\left(\theta|u_k\right)\right)\right)g\left(u_k\,|\,0,\sigma_{u_k}^2\right)du_k\right]\right\}$$
$$p\left(r_k|\theta\right) \tag{7}$$

under the Rasch testlet model with binary data, where $p\left(r_k|\theta\right)=\int p\left(r_k|\theta,u_k\right)g\left(u_k\,|\,0,\ \sigma_{u_k}^2\right)du_k$ is the probability of getting a sum score of r_k from testlet k after marginalizing out the nuisance dimension. The calculation of $p\left(r_k|\theta,u_k\right)$ is described later in this section, where it was carried out by using the Lord-Wingersky algorithm (Lord & Wingersky, 1984) .

On the other hand, the variance of the loglikelihood can also be computed for each testlet and summed up as follows:

$$Var\left(l\left(\theta|\mathbf{y}\right)\right)=\sum_{k=1}^{K}\left[Var\left(l\left(\theta\,|\,r_k\right)\right)\right]=\sum_{k=1}^{K}\left[E\left(l^2\left(\theta\,|\,r_k\right)\right)-\left(E\left(l\left(\theta\,|\,r_k\right)\right)\right)^2\right].$$

Let $\boldsymbol{y}_k$ denote the vector of item scores for testlet k, and $\mathbb{y}_{r_k}$ denote the set of score patterns that leads to a sum score of r_k. In the equation above

$$E\left(l^2\left(\theta \mid r_k\right)\right)=\sum_{r_k=0}^{n_k}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}} p\left(\boldsymbol{y}_k|\theta\right)\left\{\left(\sum\nolimits_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right.\right.$$
$$\left.\left.+\log\left[\int Exp\left(u_k r_k+\sum_{j=1}^{n_k}\log\left(q_{jk}\left(\theta|u_k\right)\right)\right) g\left(u_k \mid 0,\sigma_{u_k}^2\right) du_k\right]\right)^2\right\}, \tag{8}$$

where $p\left(\boldsymbol{y}_k|\theta\right)$ is the probability of getting a score pattern $\boldsymbol{y}_k$ after marginalizing out the nuisance dimension. The computational burden of the above formula is driven by the number of possible score pattern (2^{n_k}) and can become substantial when n_k is large. Therefore, we offer a workaround which is based on the Lord-Wingersky algorithm.

Setting

$$L_{r_k}=\log\left[\int \mathrm{Exp}\left(u_k r_k+\sum_{j=1}^{n_k}\log\left(q_{jk}\left(\theta|u_k\right)\right)\right) g\left(u_k \mid 0,\sigma_{u_k}^2\right) du_k\right],$$

we can rewrite (5) as follows:

$$E\left(l^2\left(\theta \mid r_k\right)\right)=\int\sum_{r_k=0}^{n_k}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_{jk}}\right)\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)+L_{r_k}\right)^2 g\left(u_k \mid 0,\sigma_{u_k}^2\right) du_k.$$

Rewrite

$$\begin{aligned}
&\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_{jk}}\right)\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)+L_{r_k}\right)^2\\
&=\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_{jk}}\right)\left(\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right)^2+2\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right)L_{r_k}+L_{r_k}^2\right)\\
&=\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_{jk}}\right)\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right)^2+2L_{r_k}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_j}\right)\left(\sum_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right)\\
&\quad+L_{r_k}^2 p\left(r_k|\theta\right),
\end{aligned}$$

and define

$$W_m\left(n_k,r_k\right)=\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(\prod_{j=1}^{n_k} p_{y_{jk}}\right)\left(\sum\nolimits_{j=1}^{n_k} y_{jk}\left(\theta-b_{jk}\right)\right)^m$$

if $0 \leq r_k \leq n_k$ and otherwise, we now have

$$E\left(l^2\left(\theta \mid r_k\right)\right) = \int \sum_{r_k=0}^{n_k} \left(W_2\left(n_k, r_k\right) + 2L_{r_k} W_1\left(n_k, r_k\right) + L_{r_k}^2 W_0(n_k, r_k)\right) g\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k. \tag{9}$$

For $m = 0$, $W_0\left(n_k, r_k\right)$ is the probability of obtaining a sum score of r_k for a testlet with n_k items and can be computed recursively using the Lord-Wingersky algorithm. For simplicity, let $p_{jk} = p_{jk}\left(\theta|u_k\right)$ and $q_{jk} = q_{jk}\left(\theta|u_k\right)$.

For $n_k = 1$,

$$W_0\left(1, 1\right) = p_{jk},$$
$$W_0\left(1, 0\right) = q_{jk}.$$

For $n_k = 2,\ 3,\ 4, \cdots$

$$W_0\left(n_k, r_k\right) = q_{n_k k} W_0\left(n_k - 1, r_k\right) + p_{n_k k} W_0\left(n_k - 1, r_k - 1\right).$$

Similarly, we can extend the Lord-Wingersky algorithm to compute $W_1\left(n_k, r_k\right)$ and $W_2\left(n_k, r_k\right)$ recursively. For testlet k, let $\mathbf{y}_k^{'}$ denote the vector of the first $n_k - 1$ item scores, and $\mathbb{y}_{r_k}^{'}$ denote the set of score patterns for the first $n_k - 1$ items that lead to a sum score of r_k. Then

$$\begin{aligned}
W_1(n_k, r_k) &= q_{n_k} \sum_{\mathbf{y}_k^{'} \in \mathbb{y}_{r_k}^{'}} \left(\prod_{j=1}^{n_k-1} p_{y_{jk}}\right) \left(\sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right) \\
&\quad + p_{n_k} \sum_{\mathbf{y}_k^{'} \in \mathbb{y}_{r_k-1}^{'}} \left(\prod_{j=1}^{n_k-1} p_{y_{jk}}\right) \left(\left(\theta - b_{n_k k}\right) + \sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right) \\
&= q_{n_k} W_1\left(n_k - 1, r_k\right) + \left(\theta - b_{n_k k}\right) p_{n_k} W_0\left(n_k - 1, r_k - 1\right) + p_{n_k} W_1\left(n_k - 1, r_k - 1\right). \\
W_2(n_k, r_k) &= \sum_{\mathbf{y}_k^{'} \in \mathbb{y}_{r_k}^{'}} \left(\prod_{j=1}^{n_k} p_{y_{jk}}\right) \left(y_{n_k k}\left(\theta - b_{n_k k}\right)^2 + 2y_{n_k k}\left(\theta - b_{n_k k}\right) \sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right. \\
&\quad \left. + \left(\sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right)^2\right) \\
&= q_{n_k} \sum_{\mathbf{y}_k^{'} \in \mathbb{y}_{r_k}^{'}} \left(\prod_{j=1}^{n_k-1} p_{y_{jk}}\right) \left(\sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right)^2 \\
&\quad + p_{n_k} \sum_{\mathbf{y}_k^{'} \in \mathbb{y}_{r_k-1}^{'}} \left(\prod_{j=1}^{n_k-1} p_{y_{jk}}\right) \left(\left(\theta - b_{n_k k}\right)^2 + 2\left(\theta - b_{n_k k}\right) \sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right. \\
&\quad \left. + \left(\sum_{j=1}^{n_k-1} y_{jk}\left(\theta - b_{jk}\right)\right)^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= q_{n_k} W_2\left(n_k-1, r_k\right)+\left(\theta-b_{n_k k}\right)^2 p_{n_k} W_0\left(n_k-1, r_k-1\right) \\
&\quad+2\left(\theta-b_{n_k k}\right) p_{n_k} W_1\left(n_k-1, r_k-1\right) \\
&\quad+p_{n_k} W_2\left(n_k-1, r_k-1\right).
\end{aligned}
$$

This extended version of the Lord-Wingersky algorithm significantly reduced the computational burden $E\left(l^2\left(\theta \mid r_k\right)\right)$. Also, note that $W_0\left(n_k, r_k\right) = \sum_{\boldsymbol{y}_k \in \boldsymbol{y}_{r_k}} \left(\prod_{j=1}^{n_k} p_{y_{jk}}\right) = p\left(r_k|\theta, u_k\right)$. By marginalizing out u_k as follows,

$$
p\left(r_k|\theta\right) = \int p\left(r_k|\theta, u_k\right) g\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k,
$$

we obtain $p\left(r_k|\hat{\theta}\right)$. This is the marginal probability of summed score needed in the computation of Eq. (7). To this point, all the components for computing $l_{zt}(\theta)$ have been derived.

3.2. *Variance Correction of* l_{zt}

Define the numerator of l_{zt} as $h\left(\theta|\boldsymbol{y}\right)$. When MMLE $\hat{\theta}$ is used,

$$
h\left(\hat{\theta}|\boldsymbol{y}\right) = l\left(\hat{\theta}|\boldsymbol{y}\right) - E\left(l\left(\hat{\theta}|\boldsymbol{y}\right)\right).
$$

Based on the Taylor series expansion for $\hat{\theta}$ around θ,

$$
h\left(\hat{\theta}|\boldsymbol{y}\right) = h\left(\theta|\boldsymbol{y}\right) + h'\left(\theta|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right) + r\left(\hat{\theta}\right),
$$

where $r\left(\hat{\theta}\right)$ is the remainder. In Appendix B, we prove that this remainder is negligible. Therefore, asymptotically

$$
\frac{1}{\sqrt{K}} h\left(\hat{\theta}|\boldsymbol{y}\right) = \frac{1}{\sqrt{K}} h\left(\theta|\boldsymbol{y}\right) + \frac{1}{\sqrt{K}} h'\left(\theta|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right)
$$

or

$$
\frac{1}{\sqrt{K}} h\left(\theta|\boldsymbol{y}\right) = \frac{1}{\sqrt{K}} h\left(\hat{\theta}|\boldsymbol{y}\right) - \frac{1}{\sqrt{K}} h'\left(\theta|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right),
$$

when $K \to \infty$. $\frac{1}{\sqrt{K}} h\left(\theta|\boldsymbol{y}\right)$ is asymptotically normal with mean of 0 and variance given by

$$
Var\left(\frac{1}{\sqrt{K}} h\left(\hat{\theta}|\boldsymbol{y}\right)\right) + \frac{1}{K} h'\left(\theta|\boldsymbol{y}\right)^2 Var\left(\hat{\theta}-\theta\right) + \frac{1}{K} h'\left(\theta|\boldsymbol{y}\right) * Cov\left(h\left(\hat{\theta}|\boldsymbol{y}\right), \left(\hat{\theta}-\theta\right)\right).
$$

A side product of the simulation studies presented in the next section is an investigation of the magnitude of the covariance term above. In a nutshell, at each true θ values of {-2, -1, 0, 1, 2}, 10000 test cases were simulated and the correlations between $h\left(\hat{\theta}|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right)$ were computed

for the Rasch testlet model as well as the unidimensional Rasch model. The results, which are presented in Appendix C, indicated that the covariance term is generally very close to 0. Therefore, by omitting the covariance term, we have that the sampling variation of $\frac{1}{\sqrt{K}}h\left(\theta|\mathbf{y}\right)$ is bigger than the sampling variation of $\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)$ by the term of $\frac{1}{K}h'\left(\theta|\mathbf{y}\right)^2 Var\left(\hat{\theta}-\theta\right)$, or in other words, $\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)$ is asymptotically normal with mean 0 and variance $Var\left(\frac{1}{\sqrt{K}}h\left(\theta|\mathbf{y}\right)\right) - \frac{1}{K}h'\left(\theta|\mathbf{y}\right)^2 Var\left(\hat{\theta}-\theta\right)$. The denominator used for normalization of $\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)$ is estimated by the point estimate of the variance, which has the same value asymptotically when replacing θ by $\hat{\theta}$. That is, the variance of $\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)$ is estimated by $Var\left(\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)\right) - \frac{1}{K}h'\left(\hat{\theta}|\mathbf{y}\right)^2 Var\left(\hat{\theta}-\theta\right)$. So eventually

$$\frac{\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)}{\sqrt{Var\left(\frac{1}{\sqrt{K}}h\left(\hat{\theta}|\mathbf{y}\right)\right) - \frac{1}{K}h'\left(\hat{\theta}|\mathbf{y}\right)^2 Var\left(\hat{\theta}-\theta\right)}}$$

is asymptotically standard normal. Note that $Var\left(\hat{\theta}-\theta\right)$ is in fact the inverse of the expected Fisher information provided by all the items in the test, or in another term, the inverse of the test information. Thus, we can consequently define the new person fit z-statistics as

$$l_{zt}^{*}\left(\hat{\theta}\right) = \frac{h\left(\hat{\theta}|\mathbf{y}\right)}{\sqrt{Var\left(l\left(\hat{\theta}|\mathbf{y}\right)\right) - \frac{\left(h'\left(\hat{\theta}|\mathbf{y}\right)\right)^2}{I\left(\hat{\theta}\right)}}}. \tag{10}$$

where $I\left(\hat{\theta}\right)$ is the test information at $\theta = \hat{\theta}$, defined as

$$I\left(\hat{\theta}\right) = \sum_{k=1}^{K}\left\{\sum_{r_k=0}^{n_k}\left[\left(\frac{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log\left(q_{jk}\left(\hat{\theta}|u_k\right)\right)\right)\left(r_k - \sum_{j}^{n_k} p_{jk}\left(\hat{\theta}|u_k\right)\right) g\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k}{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log\left(q_{jk}\left(\hat{\theta}|u_k\right)\right)\right) g\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k}\right)^2 p\left(r_k|\hat{\theta}\right)\right]\right\}.$$

Appendix D shows how $I\left(\hat{\theta}\right)$ was derived. It can now be recognized that the variance correction applied here for the Rasch testlet model with an MMLE ability estimate has the same form as what was shown earlier (in the review of l_z and l_z^* section) for the unidimensional model when MLE is used. Naturally, l_{zt} and l_{zt}^* reduce to l_z and l_z^*, respectively when no cluster effect is present. To compute $h'\left(\hat{\theta}|\mathbf{y}\right)$ in Eq. (10), note that with MMLE

$$h'\left(\hat{\theta}|\mathbf{y}\right) = 0 - \frac{dE\left(l\left(\hat{\theta}|\mathbf{y}\right)\right)}{d\hat{\theta}} = -\sum_{k=1}^{K}\frac{dE\left(l\left(\hat{\theta}|r_k\right)\right)}{d\hat{\theta}}. \tag{11}$$

Based on Eq. (7), $\frac{dE\left(l\left(\hat{\theta}|r_k\right)\right)}{d\hat{\theta}}$ is computed as

$$\frac{dE\left(l\left(\hat{\theta}|r_k\right)\right)}{d\hat{\theta}}$$

$$= \int \left(\sum_{j=1}^{n_k} \left(p'_{jk}\left(\hat{\theta}|u_k\right)\left(\hat{\theta}-b_{jk}\right) + p_{jk}\left(\hat{\theta}|u_k\right)\right)\right) N\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k$$

$$+ \sum_{r_k=0}^{R_k} \left\{ \log\left[\int \text{Exp}\left(r_k u_k + \sum_{j=1}^{n_k} \log\left(q_{jk}\left(\hat{\theta}|u_k\right)\right)\right) N\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k \right] \right\} p'\left(r_k|\hat{\theta}\right)$$

$$+ \sum_{r_k=0}^{R_k} \frac{\int \text{Exp}\left(r_k u_k + \sum_{j=1}^{n_k} \log\left(q_{jk}\left(\hat{\theta}|u_k\right)\right)\right)\left(\sum_{j=1}^{n_k} \frac{q'_{jk}\left(\hat{\theta}|u_k\right)}{q_{jk}\left(\hat{\theta}|u_k\right)}\right) N\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k}{\int \text{Exp}\left(r_k u_k + \sum_{j=1}^{n_k} \log\left(q_{jk}\left(\hat{\theta}|u_k\right)\right)\right) N\left(u_k \mid 0, \sigma_{u_k}^2\right) du_k}$$

$$p\left(r_k|\hat{\theta}\right).$$

The only unknown in the above equation is $p'\left(r_k|\hat{\theta}\right)$, i.e., the derivative of $p\left(r_k|\hat{\theta}\right)$ with respect to $\hat{\theta}$. While $p\left(r_k|\hat{\theta}\right)$ is computed recursively by our extended Lord-Wingersky algorithm, $p'\left(r_k|\hat{\theta}\right)$ can also be computed recursively as follows by applying the product rule to $W_0(n_k, r_k)$:

for $n_k = 1$,

$$W'_0(1,1) = p'_{jk},$$
$$W'_0(1,0) = q'_{jk},$$

and for $n_k = 2,\ 3,\ 4, \cdots$

$$W'_0(n_k, r_k) = q'_{n_k k} W_0(n_k - 1, r_k) + q_{n_k k} W'_0(n_k - 1, r_k) + p'_{n_k k} W_0(n_k - 1, r_k - 1)$$
$$+ p_{n_k k} W'_0(n_k - 1, r_k - 1).$$

Therefore,

$$p'\left(r_k|\hat{\theta}\right) = \int p'\left(r_k|\hat{\theta}, u_k\right) N\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k = \int W'_0(n_k, r_k)\, N\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k.$$

To this point, all components to compute $l_{zt}^*\left(\hat{\theta}\right)$ have been derived.

4. Simulation Study

4.1. Type I Error Rates

This section presents the results of a simulation study conducted to investigate the empirical type I error rate of l_{zt} (i.e., before correction) and l_{zt}^* (i.e., after correction). Items used in the studies were sampled from an operational item bank of a K-12 standardized assessment in the United States. The test length varied at 6 testlets and 12 testlets. Table 1 presents the summary of items.

All items have been previously calibrated, and their parameters were taken as fixed values. For each test length condition, true θ values from -2 to 2 with a step of 1 were selected, and 10,000 simulated test datasets were generated at each θ value. $\hat{\theta}$s were then estimated by the MMLE in each test and used for the calculation of person fit statistics. Critical values were

TABLE 1.
Summary of Items Used in the Simulations.

Number of testlets	6	12
Total number of items	43	96
b parameter range	[−1.96, 1.93]	[−1.96, 2.00]
b parameter mean	0.07	−0.22
b parameter standard deviation	0.99	0.80
σ_u^2 range	[0.003, 0.629]	[0.157, 1.253]
σ_u^2 mean	0.28	0.55
σ_u^2 standard deviation	0.22	0.37

chosen corresponding to nominal error rates of $\alpha = .05$ and .01 to identify aberrant responses. Occasionally, there were cases where all items were answered correctly or incorrectly. Since these cases do not provide information on how the IRT model fit to the data as the MMLE is not defined (i.e., $\hat{\theta}$ is ∞ or $-\infty$), they were discarded when summarizing the simulation results. The highest value of the discard rate at any given θ was 0.001 with the 6-testlet test when $\theta = -2$. In addition, to provide a baseline for comparison, instead of using $\hat{\theta}$, l_{zt} was also computed for the simulated responses by plugging in the true θ.

Figure 1 shows the kernel density of l_{zt} and l_{zt}^* (both computed with $\hat{\theta}$) overlaid with the standard normal distribution for each condition. When $\theta = 0$, both l_{zt} and l_{zt}^* were close to a standard normal distribution. However, as θ became more extreme, the variance of l_{zt} diminished and the distribution of l_{zt} deviated from standard normal, whereas l_{zt}^* remained close to standard normal. Consequently, as shown in Table 2, Type I error rates of l_{zt} computed with $\hat{\theta}$ were reasonably close to the nominal rate at $\theta = 0$, but were much smaller when θ became more extreme. On the contrary, the values of l_{zt}^* were always close to the nominal rates and were often substantially better than those of l_{zt}. It was also found that the baseline Type I error rates of l_{zt} computed with θ (rows denoted with "true θ" in the table) were somewhat higher than the nominal rates, especially when θ became more extreme. l_{zt}^*, which was computed with $\hat{\theta}$, provided Type I error rates that are closer to the nominal rate even when compared to baseline rates provided by l_{zt} computed with θ. Finally, the asymptotic approximation of l_{zt}^* became better when test-length increased as one would expect.

4.2. Power

To investigate the power of l_{zt}^*, the data used in the investigation of the Type I error rate were manipulated to reflect aberrant responses. A spuriously-high-score scenario was created where 10% (or 30%) of the most difficult items among the test were assigned responses of 1, and a spuriously-low-score scenario was created where 10% (or 30%) of the easiest items were assigned responses of 0. Similar to what was done in the Type I error rate analysis, cases where the MMLE was not defined were discarded. The highest value of the discard rate at any given θ was 0.069 with the 6-testlet test when $\theta = -2$ and the data has 30% aberrantly low scores. The overall discard rate across all conditions was 0.003. Tables 3 and 4 indicate that at θ values where aberrant responses are more likely to arise (i.e., low θ values for the spuriously-high-score scenario and high θ values for the spuriously-low-score scenario), l_{zt}^* offered sufficiently large power of detection. Although l_{zt} also offered more power at those θ values than other values, the power of l_{zt}^* was always higher than that of l_{zt}. For a relatively short test with relatively less aberrant responses, l_{zt} lacked its power even at θ values where aberrant responses are more likely to arise, whereas l_{zt}^* offered decent power. As expected, the power increased as test length and the percentage of aberrant responses increased.

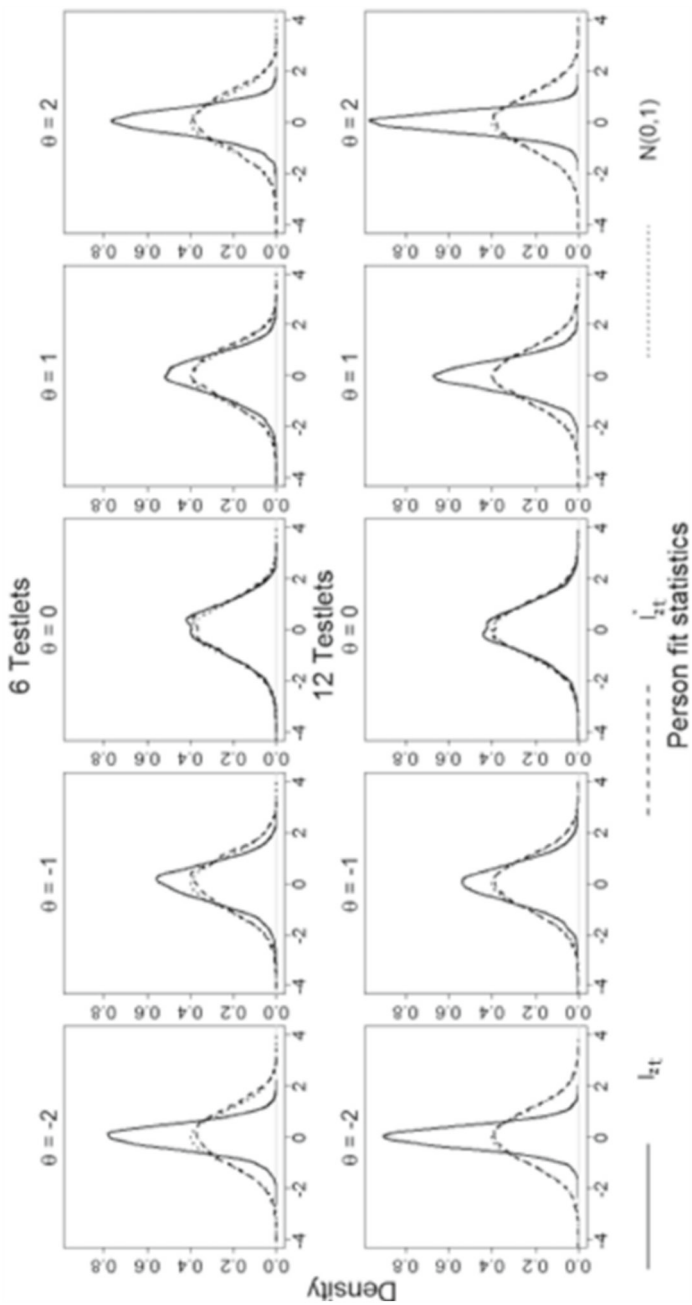

FIGURE 1.
Distributions of l_{zt} and l_{zt}^{*} overlaid with the standard normal distribution for each condition in the simulation study.

TABLE 2.
Type I Error Rate from Simulation.

θ	−2	−1	0	1	2
6 Testlets					
l_{zt} ($\alpha = 0.05$, true θ)	0.082	0.069	0.062	0.071	0.087
l_{zt} ($\alpha = 0.05$)	0.002	0.021	0.048	0.027	0.004
l^*_{zt} ($\alpha = 0.05$)	0.057	0.057	0.060	0.059	0.057
l_{zt} ($\alpha = 0.01$, true θ)	0.027	0.021	0.016	0.021	0.031
l_{zt} ($\alpha = 0.01$)	0.000	0.004	0.013	0.005	0.000
l^*_{zt} ($\alpha = 0.01$)	0.015	0.016	0.016	0.016	0.015
12 Testlets					
l_{zt} ($\alpha = 0.05$, true θ)	0.073	.060	0.051	.067	0.075
l_{zt} ($\alpha = 0.05$)	0.001	0.021	0.041	0.007	0.000
l^*_{zt} ($\alpha = 0.05$)	0.053	0.057	0.054	0.058	0.053
l_{zt} ($\alpha = 0.01$, true θ)	0.023	0.013	0.011	0.016	0.022
l_{zt} ($\alpha = 0.01$)	0.000	0.002	0.009	0.000	0.000
l^*_{zt} ($\alpha = 0.01$)	0.012	0.013	0.013	0.014	0.013

TABLE 3.
Power Under Spuriously-high-score Scenario.

Aberrant response rate	10%					30%				
θ	−2	−1	0	1	2	−2	−1	0	1	2
6 Testlets										
l_{zt} ($\alpha = 0.05$)	0.88	0.83	0.66	0.23	0.02	1.00	1.00	0.90	0.38	0.04
l^*_{zt} ($\alpha = 0.05$)	0.97	0.88	0.72	0.50	0.30	1.00	1.00	1.00	0.93	0.64
l_{zt} ($\alpha = 0.01$)	0.50	0.55	0.39	0.07	0.00	1.00	0.99	0.71	0.12	0.00
l^*_{zt} ($\alpha = 0.01$)	0.99	0.89	0.69	0.46	0.24	1.00	1.00	0.98	0.75	0.35
12 Testlets										
l_{zt} ($\alpha = .05$)	1.00	1.00	0.85	0.22	0.01	1.00	1.00	0.96	0.37	0.01
l^*_{zt} ($\alpha = 0.05$)	1.00	1.00	0.96	0.81	0.51	1.00	1.00	1.00	0.99	0.84
l_{zt} ($\alpha = 0.01$)	0.96	0.96	0.58	0.04	0.00	1.00	1.00	0.76	0.07	0.00
l^*_{zt} ($\alpha = 0.01$)	1.00	0.98	0.85	0.53	0.23	1.00	1.00	1.00	0.94	0.58

TABLE 4.
Power Under Spuriously-low-score Scenario.

Aberrant response rate	10%					30%				
θ	−2	−1	0	1	2	−2	−1	0	1	2
6 Testlets										
l_{zt} ($\alpha = 0.05$)	0.01	0.15	0.61	0.87	0.90	0.03	0.34	0.90	1.00	1.00
l^*_{zt} ($\alpha = 0.05$)	0.24	0.43	0.71	0.90	0.98	0.69	0.95	1.00	1.00	1.00
l_{zt} ($\alpha = 0.01$)	0.00	0.03	0.33	0.60	0.53	0.00	0.10	0.69	0.99	1.00
l^*_{zt} ($\alpha = .01$)	0.08	0.18	0.42	0.71	0.91	0.37	0.77	0.98	1.00	1.00
12 Testlets										
l_{zt} ($\alpha = 0.05$)	0.01	0.28	0.84	0.96	0.98	0.02	0.48	0.98	1.00	1.00
l^*_{zt} ($\alpha = 0.05$)	0.41	0.66	0.89	0.98	1.00	0.81	0.98	1.00	1.00	1.00
l_{zt} ($\alpha = 0.01$)	0.00	0.08	0.62	0.76	0.63	0.00	0.16	0.89	1.00	1.00
l^*_{zt} ($\alpha = 0.01$)	0.17	0.39	0.70	0.92	0.99	0.58	0.91	1.00	1.00	1.00

5. Application to Real Data

An advantage of l_{zt}^* is that it allows person fit evaluation for not only tests consist of items modeled by either unidimensional models or the Rasch testlet model, but also for novel tests that are modeled by a mixture of these two types of components. This section demonstrates such an application of l_{zt}^* to a U.S. statewide test assessing the Next Generation of Science Standards (NGSS). The test is mainly comprised of item clusters. An item cluster represents a series of interrelated examinee interactions directed toward describing, explaining and predicting scientific phenomena. Within each item cluster, a set of explicit assertions were made about examinee's knowledge or skills according to specific features they've demonstrated through their interactions with the item cluster. In this setting, an *assertion* is analogous to a traditional item, and it was scored as 1 if it was asserted and 0 if it is not asserted. An *item cluster* is an item bundle (testlet) consists of multiple assertions. To account for the conditional dependency amount assertions within an item cluster, the part of the latent structure that describes the item clusters is the same as the Rasch testlet model. That is, an overall science dimension as well as additional "nuisance" dimensions corresponding to the bundling of the items. On the other hand, the model also allows a subset of assertions to depend only on the overall science dimension. These so-called *stand-alone assertions* typically pertain to shorter items (typically less than 4 assertions within an item) and were assumed independent given the overall dimension. This part of the latent structure is the same as the unidimensional Rasch model. Figure 2 shows the model graphically.

The item pool of the assessment consisted of 27 item clusters and 24 stand-alone items. The test was administered online using a linear-on-the-fly test design (LOFT) such that each examinee received 6 item clusters and 12 stand-alone items at random that meets the test blueprint. A total of 12,026 examinees who completed all 18 items were included in the analysis. All the items have been previously calibrated. Table 5 presents a summary of the 18 test items an individual would typically receive.

MMLE was used to estimate examinee abilities. Since no examinee answered all items correctly or all items incorrectly, no MMLE estimate was undefined. l_{zt}^* values were computed for every examinee. Specifically, using the general definition of l_{zt}^* in Eq. (10), each component involved in the computation can be calculated separately for the item clusters and for the stand-alone assertions, and then simply combined (added) to produce the statistics. Examinees were flagged if their l_{zt}^* values were below the critical value of the nominal error rates at $\alpha = .05$, and further flagged if below the critical value at $\alpha = .01$. Three examinee groups were then created based on the flags: No Flag/ Flagged at .05/ Flagged at .01. Within each subset, a "person-total" correlation is computed for each examinee. Analogous to the item-total correlation, the person-total correlation is essentially the correlation between an examinee's item scores and the average scores of the same items by all examinees. One would expect an examinee to be more likely to fit if his/her item scores agree better with the item scores of other examinees, and therefore less likely to be flagged. The person-total correlation was computed at both item level and the assertion level, where at the assertion level the assertion scores were used, and at the item level the average assertion scores within an item were used. For both levels, the person-total correlation was averaged within each examinee subsets. In addition, the l_z^* statistics were computed for the same examinees with MLE ability while ignoring the cluster effect. The same procedure of flagging and person-total correlation computation described above were applied. Table 6 presents results for both l_{zt}^* and l_z^*. Both methods yielded similar correlations for the group without flag. However, as expected for l_{zt}^*, the group with no flag has correlations much higher than the flagged groups at both the item and assertion levels, and the lowest correlations were observed with the group flagged at .01. On the other hand, for l_z^*, both flagged groups have relative high correlations that are close to the ones found in the group with no flag.

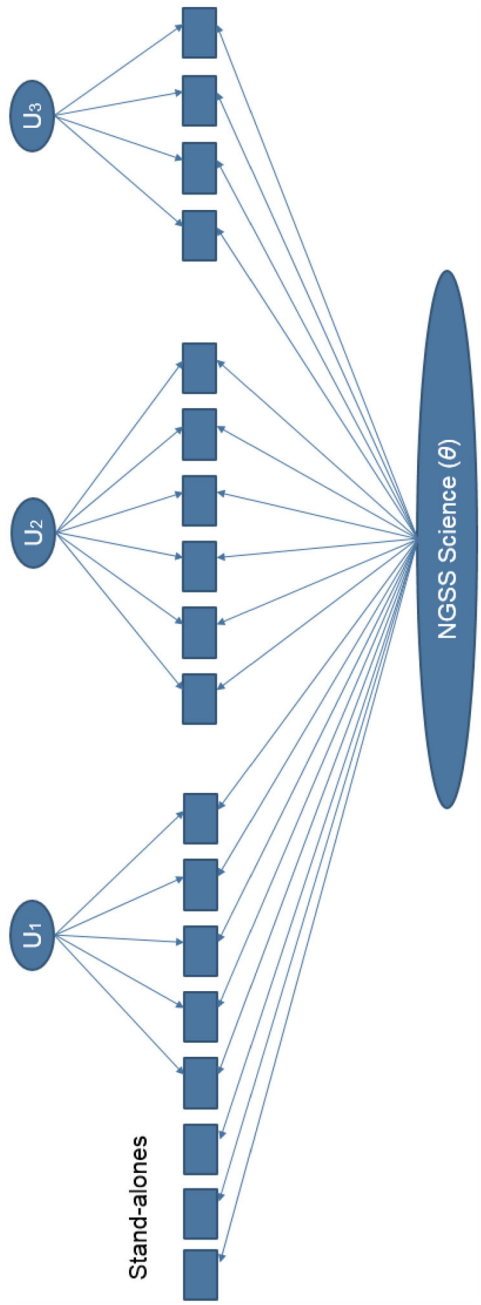

FIGURE 2.
Directed graph of the IRT model in the real data analysis.

TABLE 5.
Summary of Items Parameters for an 18-item Test, averaged over all examinees.

Item type	12 Stand-alones	6 Clusters	All 18 items
b parameter range	[−1.57, 2.57]	[−2.78, 2.89]	[−2.81, 2.95]
b parameter mean	−0.06	0.17	0.12
b parameter standard deviation	0.97	1.08	1.06
σ_u^2 range	NA	[0.17, 1.85]	NA
σ_u^2 mean	NA	0.63	NA
σ_u^2 standard deviation	NA	0.56	NA

TABLE 6.
Average Person-total Correlation among Examinee Subsets.

	Examinee subset	Average person-total correlation	
		Assertion level	Item level
l_{zt}^*	No Flag	0.41	0.41
	Flagged at 0.05	0.17	0.15
	Flagged at 0.01	0.12	0.10
l_z^*	No Flag	0.43	0.44
	Flagged at 0.05	0.34	0.33
	Flagged at 0.01	0.33	0.32

l_z^* is computed when cluster effects were ignored

For each examinee within a subset, a few more detail can be depicted to examine the agreement among the p-value of l_{zt}^*, person-total correlation, and the pattern of item scores. First, the assertions an examinee received were grouped. The 6 item clusters, together with all the stand-alone assertions naturally formed a total of 7 groups. These groups of assertions were then arranged in a descending order by the average assertion difficulty. The average assertion scores of each group were calculate for the examinee and plotted against the grouping. Figure 3 shows the plots for four examinees. The title of a panel shows the person-total correlation and the p-value of l_{zt}^*, respectively for an examinee. The examinee on the top-left panel is from the subset with no flags. In general, this examinee's average item group scores increased as the difficulty of the item group decreased (except for one obvious outlier) and therefore had a moderately high person-total correlation of 0.29. This examinee was not classified as a misfit with a p-value of 0.579. The examinee on the top-right panel is also from the subset with no flags. A strong increasing pattern was observed. This examinee had a correlation of 0.77 and was not classified as misfit with a higher p-value of 0.967 than the examinee on top-left. On the contrary, examinees in the bottom panels are from the flagged subsets. The examinee on the bottom-left had a p-value of 0.028, and a low correlation of 0.1. The pattern of the average item group scores against average item group difficulty seemed to be random. Finally, the examinee on the bottom-right had a p-value of 0.002. A decreasing pattern and a slightly negative person total correlation of −0.07 was observed. These figures suggested the flagging of the l_{zt}^* agreed with other sources of evidence when assessing the fit for the same person.

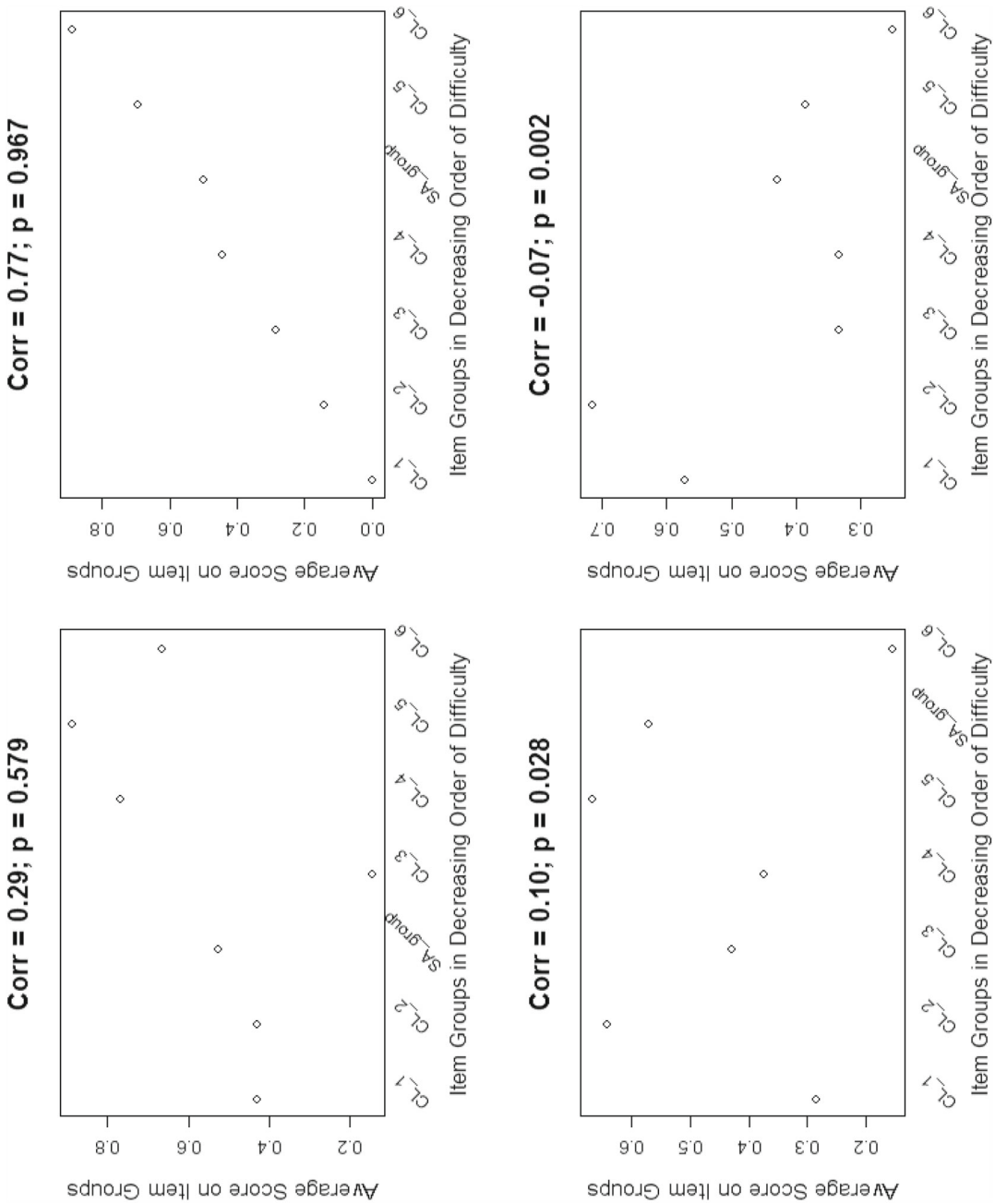

FIGURE 3.
Agreement among the p-value of $l_{z^*}^*$, person-total correlation, and the pattern of item scores for four examinees in the real data analysis.

6. Conclusion and Discussion

IRT testlet models have been frequently put into practice where the latent trait corresponding to the overall dimensional is usually of primary interest while other dimensions are incorporated as nuisance dimensions only to address the local dependencies between items within clusters. Moreover, unlike a traditional test which usually assumes either a unidimensional or multidimensional latent structure for every item in the test, novel tests (and models) may incorporate both components in their latent structure. Just like it is with unidimensional models, person fit evaluation with these models is an important part of the model-data fit evaluation that facilitates the delivery of reliable and valid test results. However, research on person fit statistics beyond unidimensional model is relatively scarce. The current study embarks on an effort to fill this gap by offering a person fit z-statistics appropriate for the Rasch testlet model, traditional unidimensional models, as well as models that have both components. Under the Rasch testlet model, the proposed person fit indices, l_{zt} and its corrected version l_{zt}^*, are extensions of the well-known existing indices l_z and l_z^* for unidimensional models. In the simulation study, the Type I error rate and power of the new statistics under the Rasch testlet model were investigated and found to be consistent with the results of their counterparts under unidimensional models in the literature (Sinharay, 2016; Snijders, 2001) . l_{zt}^* provided close to nominal Type I error rate and good power to detect aberrant response. Furthermore, this method of extension entailed a generalized approach to correct the variance of the loglikelihood when maximum likelihood estimation was used to estimate ability parameters. Under traditional unidimensional model, l_{zt} and l_{zt}^* reduce to l_z and l_z^*, respectively. This generalization keeps person fit evaluation with both the unidimensional models and the Rasch testlet model under the same framework and allows for person fit evaluation with models that have both components in their latent structure. The real data analysis example shows the utility of l_{zt}^* under such a circumstance, which is otherwise not possible with l_z^* without violating the original model assumption.

While developing l_{zt} and l_{zt}^* for their use with Rach testlet model, the Lord-Wingersky algorithm was extended in a few ways to achieve efficient computation. These extensions are considered another important contribution of this article. In a nutshell, three kinds of extensions of the algorithm were presented. First, realizing the fact that the expected value of the entire data loglikelihood under Rasch testlet model can be accumulated testlet by testlet using within testlet sum score loglikelihoods, the Lord-Wingersky algorithm was extended. Note that this straightforward extension is the same as what was described by Cai (2015) in his Equation 16 or 20, which took advantage of the assumed bifactor structure (or more generally, the two-tier structure). Second, the Lord-Wingersky algorithm is further extended for computing components in the variance of the loglikelihood. An implication of this extension is that not only can one use the algorithm to compute the probabilities of sum scores (e.g., $W_0(n_k, r_k)$), but one can also define other related quantities (e.g., $W_1(n_k, r_k)$ and $W_2(n_k, r_k)$) to make use of the recursive nature of the algorithm as needed. The third extension of the Lord-Wingersky algorithm was applied when computing $\frac{dE\left(l\left(\hat{\theta}|r_k\right)\right)}{d\hat{\theta}}$, where the derivative of the sum score within a testlet was needed. Although the extension was again a straightforward application of the product rule from basic calculus, it avoided doing numerical integration directly on $\frac{dE\left(l\left(\hat{\theta}|r_k\right)\right)}{d\hat{\theta}}$, and therefore increased the accuracy of the results as well as the speed of computations.

Like most of the person fit statistics in the literature, l_{zt}^* is a statistic pertaining to one individual. A statistically significant l_{zt}^* does not necessarily mean an examinee had abnormal testing behaviors. Further investigation of the flagged examinees must be conducted, especially when drawing high-stake conclusions such as whether an examinee cheated during the test. Nonetheless, it can serve as a screening mechanism to find individuals with potential testing behavior related

issues. How liberal/rigid the screening criteria is would depend on resource available. When an aggregated unit of examinees is of concern, person fit statistics like l_{zt}^* can also be useful by either simply checking the percentage of examinees flagged within the aggregated unit or constructing t statistics to flag units statistically.

One limitation of l_{zt}^* is that the current extension only concerns the Rasch testlet model when θ is estimated by MMLE. The relatively straightforward derivation of l_{zt}^* relied on the fact that a sufficient statistic exists for a given testlet, as well as the fact that the nuisance dimension is marginalized out in MMLE. There could be scenarios where people prefer to use a more complex model such as a bifactor model not belonging to the Rasch family or other multidimensional IRT models where latent trait on multiple dimensions are of interest. There could also be scenarios where EAP, MLE, or MAP (maximum-a-posteriori) estimators are preferred. Under those scenarios, the derivation of l_{zt} and l_{zt}^* could become more challenging. In addition, there has also been a recent study that corrects the standardized person-fit statistics regarding both the use of an estimated ability and the use of a finite number of items Gorney et al. (2024). Further study is needed to explore these topics regarding the l_{zt} and l_{zt}^* statistics.

Finally, as concluded by Sinharay (2016), among others, the l_z^* statistics is appropriate when an investigator wants to test against an unspecified general and may not be the most appropriate person-fit statistic for a particular problem, such as for a computer adaptive test. Also, when item parameters are not treated as fixed but need to be estimated, any aberrant response in the data would have impact on the item parameter estimation and in turn affects the person fit statistics. As an extension of l_z^*, l_{zt}^* shares these same limitations. More research on these topics, as well as the performance of l_{zt}^* against other person fit statistics, would be helpful to practitioners.

Appendix A: Sufficient Statistic for MMLE θ under the Rasch Testlet Model

For a test consists of K testlets, the likelihood of the MMLE overall ability θ is defined as

$$
\begin{aligned}
L(\theta|\mathbf{y}) &= \prod_{k=1}^{K} \int \prod_{j=1}^{n_k} \left(p_{jk}(\theta|u_k)\right)^{y_{jk}} \left(q_{jk}(\theta|u_k)\right)^{1-y_{jk}} g\left(u_k \,|\, 0, \sigma_{u_k}^2\right) du_k \\
&= \prod_{k=1}^{K} \int \left[\prod_{j=1}^{n_k} \left(q_{jk}(\theta|u_k)\right)\right] \left[\mathrm{Exp}\left((\theta+u_k)\sum_{j=1}^{n_k} y_{jk}\right)\right] \left[Exp\left(-\sum_{j=1}^{n_k} b_{jk} y_{jk}\right)\right] \\
&\quad g\left(u_k \,|\, 0,\, \sigma_{u_k}^2\right) du_k \\
&= \prod_{k=1}^{K} \left[\mathrm{Exp}\left(-\sum_{j=1}^{n_k} b_{jk} y_{jk}\right)\right] \\
&\quad \times \prod_{k=1}^{K} \int \left[\prod_{j=1}^{n_k} \left(q_{jk}(\theta|u_k)\right)\right] \left[\mathrm{Exp}\left((\theta+u_k)\sum_{j=1}^{n_k} y_{jk}\right)\right] g\left(u_k \,|\, 0,\, \sigma_{u_k}^2\right) du_k
\end{aligned}
$$

Define $r_k = \sum_{j=1}^{n_k} y_{jk}$, we can see the that the likelihood function of θ was factored into a product of $\prod_{k=1}^{K}\left[\text{Exp}\left(-\sum_{j=1}^{n_k} b_{jk}y_{jk}\right)\right]$ which does not depend on θ and the rest of the terms which does depend on θ but only through r_k. Therefore, based on the Fisher–Neyman factorization theorem, we can conclude that vector $\{r_1, r_2, \cdots, r_K\}$ is the sufficient statistic for θ. That is, all the information about θ available in a response pattern $\boldsymbol{y}$ is given by $\{r_1, r_2, \cdots, r_K\}$.

Appendix B: Proof of the Taylor Expansion Remainder Term Being Negligible

Recall that we defined

$$h\left(\hat{\theta}|\boldsymbol{y}\right) = h\left(\theta|\boldsymbol{y}\right) + h'\left(\theta|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right) + r\left(\hat{\theta}\right)$$

By rearranging the above equation,

$$r\left(\hat{\theta}\right) = h\left(\hat{\theta}|\boldsymbol{y}\right) - h\left(\theta|\boldsymbol{y}\right) - h'\left(\theta|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right).$$

Taking derivative of $r\left(\hat{\theta}\right)$ with respect to $\hat{\theta}$ gives

$$r'\left(\hat{\theta}\right) = h'\left(\hat{\theta}|\boldsymbol{y}\right) - h'\left(\theta|\boldsymbol{y}\right).$$

According to the mean value theorem, there exists a point $\tilde{\theta}$ such that

$$h''\left(\tilde{\theta}|\boldsymbol{y}\right) = \frac{h'\left(\hat{\theta}|\boldsymbol{y}\right) - h'\left(\theta|\boldsymbol{y}\right)}{\left(\hat{\theta}-\theta\right)}.$$

Therefore,

$$r'\left(\hat{\theta}\right) = h'\left(\hat{\theta}|\boldsymbol{y}\right) - h'\left(\theta|\boldsymbol{y}\right) = h''\left(\tilde{\theta}|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right).$$

Taking the antiderivative of $r'\left(\hat{\theta}\right)$ and because $r\left(\theta\right) = 0$, we obtain

$$r\left(\hat{\theta}\right) = \frac{1}{2}h''\left(\tilde{\theta}|\boldsymbol{y}\right)\left(\hat{\theta}-\theta\right)^2.$$

We now provide some property regarding the function h.

1. $\frac{h'(\theta|\boldsymbol{y})}{K}$ is bounded for any θ

Proof. The loglikelihood function of a testlet k is defined as

$$l_k(\theta|\mathbf{y}) = \log\left[\int_{-\infty}^{\infty} \text{Exp}\left(\sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}(\theta|u_k)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}(\theta|u_k)\right)\right)\right) g\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k\right].$$

Let $t_k = G\left(u_k \mid 0,\ \sigma_{u_k}^2\right)$ be the CDF, we then have

$$l_k(\theta|\mathbf{y}) = \log\left[\int_{-\infty}^{\infty} \text{Exp}\left(\sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}(\theta|u_k)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}(\theta|u_k)\right)\right)\right) dt_k\right].$$

And since $u_k = G^{-1}\left(t_k \mid 0,\ \sigma_{u_k}^2\right)$, we have

$$l_k(\theta|\mathbf{y}) = \log\left[\int_0^1 \text{Exp}\left(\sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}\left(\theta|G^{-1}\left(t_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}\left(\theta|G^{-1}\left(t_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right)\right)\right) dt_k\right]$$

Let

$$f(t_k) = \text{Exp}\left(\sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}\left(\theta|G^{-1}\left(t_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}\left(\theta|G^{-1}\left(t_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right)\right)\right).$$

Since $f(t_k) \to 0$ when both $t_k \to 0$ and $t_k \to 1$, we can define $f(0) = 0$ and $f(1) = 0$, so that $f(t_k)$ is considered continuous in [0,1]. By applying the mean value theorem for integral, there exists a value $c_k \in (0, 1)$ and at which

$$\begin{aligned} l_k(\theta|\mathbf{y}) &= \log\left[f(c_k)\right] \\ &= \sum_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}\left(\theta|G^{-1}\left(c_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}\left(\theta|G^{-1}\left(c_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right)\right) \end{aligned}$$

Using the above equation, we see

$$l_k'(\theta|\mathbf{y}) = -\sum_{j=1}^{n_k}\left(y_{jk}q_{jk}\left(\theta|G^{-1}\left(c_k \mid 0,\ \sigma_{u_k}^2\right)\right) - \left(1-y_{jk}\right)p_{jk}\left(\theta|G^{-1}\left(c_k \mid 0,\ \sigma_{u_k}^2\right)\right)\right),$$

which can be bonded by n_k.
As for the derivatives of the expected loglikelihood, we have

$$\begin{aligned} E\left(l_k(\theta|\mathbf{y})\right) &= E\left(\sum\nolimits_{j=1}^{n_k}\left(y_{jk}\log\left(p_{jk}(\theta|d_k)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}(\theta|d_k)\right)\right)\right) \\ &= \sum_{j=1}^{n_k} E\left(y_{jk}\log\left(p_{jk}(\theta|d_k)\right) + \left(1-y_{jk}\right)\log\left(q_{jk}(\theta|d_k)\right)\right) \end{aligned}$$

$$= \sum_{j=1}^{n_k} \left(E\left(y_{jk}\right) \log\left(p_{jk}\left(\theta|d_k\right)\right) + \left(1 - E\left(y_{jk}\right)\right) \log\left(q_{jk}\left(\theta|d_k\right)\right)\right)$$

Since $E\left(y_{jk}\right) = \int \frac{\exp(\theta+u_k-b_j)}{1+\exp(\theta+u_k-b_j)} g\left(u_k \mid 0,\ \sigma_{u_k}^2\right) du_k$, using similar argument, we know there exists d_{jk} that $E\left(y_{jk}\right) = \frac{\exp(\theta+d_{jk}-b_j)}{1+\exp(\theta+d_{jk}-b_j)}$, hence

$$E\left(l_k\left(\theta|\mathbf{y}\right)\right) = \sum_{j=1}^{n_k} \left(p_{jk}\left(\theta|d_{jk}\right) \log\left(p_{jk}\left(\theta|d_k\right)\right) + \left(1 - p_{jk}\left(\theta|d_{jk}\right)\right) \log\left(q_{jk}\left(\theta|d_k\right)\right)\right)$$

and

$$\frac{dE\left(l_k\left(\theta|\mathbf{y}\right)\right)}{d\theta} = \sum_{j=1}^{n_k} \left[p_{jk}\left(\theta|d_{jk}\right) q_{jk}\left(\theta|d_{jk}\right) \log\left(p_{jk}\left(\theta|d_k\right)\right) - p_{jk}\left(\theta|d_{jk}\right) q_{jk}\left(\theta|d_{jk}\right) \log\left(q_{jk}\left(\theta|d_k\right)\right)\right]$$

Let s_1 stands for all the terms within the summation operator above. s_1 is bounded because both addends approach to 0 as θ goes to $\pm\infty$. Suppose the absolute value, $|s_1|$, is less than some real number μ_1, then $\left|\frac{dE(l_k(\theta|\mathbf{y}))}{d\theta}\right| < u_1 n_k$. Therefore,

$$\left|\frac{h'\left(\tilde{\theta}|\mathbf{y}\right)}{K}\right| < (u_1 + 1) \max\left\{n_1, n_2, \cdots, n_K\right\}.$$

This proves that $\frac{h'\left(\tilde{\theta}\right)}{K}$ is bounded for any θ.

2. $\frac{h''(\theta|\mathbf{y})}{K}$ is bounded for any θ

Proof. Using the earlier results from proof 1, we can find that

$$l_k''\left(\theta|\mathbf{y}\right) = -\sum_{j=1}^{n_k} \left(p_{jk}\left(\theta|d_k\right) q_{jk}\left(\theta|d_k\right)\right).$$

Here $p_{jk}\left(\theta|d_k\right) q_{jk}\left(\theta|d_k\right)$ is bounded by 1/4 for any value of θ, hence $l_k''\left(\theta|\mathbf{y}\right)$ is bounded by $\frac{n_k}{4}$. We also have

$$\begin{aligned}\frac{d^2E\left(l_k\left(\theta|\mathbf{y}\right)\right)}{d\theta^2} = \sum_{j=1}^{n_k} \Big[&\left(q_{jk}\left(\theta|d_{jk}\right) - p_{jk}\left(\theta|d_{jk}\right)\right) p_{jk}\left(\theta|d_{jk}\right) q_{jk}\left(\theta|d_{jk}\right) \log\left(p_{jk}\left(\theta|d_k\right)\right) \\ &+ \left(p_{jk}\left(\theta|d_{jk}\right) - q_{jk}\left(\theta|d_{jk}\right)\right) p_{jk}\left(\theta|d_{jk}\right) q_{jk}\left(\theta|d_{jk}\right) \log\left(q_{jk}\left(\theta|d_k\right)\right) \\ &+ p_{jk}\left(\theta|d_{jk}\right) q_{jk}\left(\theta|d_{jk}\right)\Big].\end{aligned}$$

Let s_2 stands for all the terms within the summation operator above. s_2 is bounded because all addends approach to 0 as θ goes to $\pm\infty$. Suppose the absolute value, $|s_2|$, is less than some real number μ_2, then $\left|\frac{d^2E(l_k(\theta|\mathbf{y}))}{d\theta^2}\right| < u_2 n_k$.

Using the above results, we show that $\left|l_k''(\theta|\mathbf{y}) - \frac{d^2E(l_k(\theta|\mathbf{y}))}{d\theta^2}\right| \leq \left(u_2 + \frac{1}{4}\right) n_k$. Therefore,

$$\left|\frac{h''(\theta)}{K}\right| = \frac{\left|\sum_{k=1}^{K}\left(l_k''(\theta|\mathbf{y}) - \frac{d^2E(l_k(\theta|\mathbf{y}))}{d\theta^2}\right)\right|}{K} \leq \frac{K\left(u_2 + \frac{1}{4}\right)\max\{n_1, n_2, \cdots, n_K\}}{K}$$
$$= \left(u_2 + \frac{1}{4}\right)\max\{n_1, n_2, \cdots, n_K\}.$$

This proves that $\frac{h''\left(\tilde{\theta}\right)}{K}$ is bounded for any θ.

Now recall that the remainder of the Taylor expansion is

$$r\left(\hat{\theta}\right) = \frac{1}{2}h''\left(\tilde{\theta}|\mathbf{y}\right)\left(\hat{\theta} - \theta\right)^2$$

For $h''\left(\tilde{\theta}|\mathbf{y}\right)\left(\hat{\theta} - \theta\right)^2$, we have

$$\frac{1}{\sqrt{K}}h''\left(\tilde{\theta}|\mathbf{y}\right)\left(\hat{\theta} - \theta\right)^2 = \sqrt{K}\left(\hat{\theta} - \theta\right)\left(\hat{\theta} - \theta\right)\frac{h''\left(\tilde{\theta}|\mathbf{y}\right)}{K}.$$

Here, $\sqrt{K}\left(\hat{\theta} - \theta\right)$ is asymptotical normal, $\left(\hat{\theta} - \theta\right)$converges to 0 in probability, and $\frac{h''\left(\tilde{\theta}|\mathbf{y}\right)}{K}$ is bounded. Therefore, $\frac{1}{\sqrt{K}}h''\left(\tilde{\theta}|\mathbf{y}\right)\left(\hat{\theta} - \theta\right)^2$ converges to 0 in probability. This indicates that the remainder is negligible.

Appendix C: Simulation Results of the Correlation Between $h\left(\hat{\theta}|y\right)$ and $\left(\hat{\theta}-\theta\right)$

True θ	Correlation between $h\left(\hat{\theta}\|y\right)$ and $\left(\hat{\theta} - \theta\right)$			
	Rasch Testlet Model		Unidimensional Rasch Model*	
	6 Testlets	12 Testlets	43 items	96 items
−2	0.0276	0.0281	0.0022	0.0073
−1	0.0091	0.0100	−0.0089	0.0012
−0	0.0168	−0.0074	−0.0074	−0.0051
−1	0.0005	−0.0213	0.0033	0.0026
−2	0.0046	−0.0327	−0.0019	−0.0126

**Note.* For the unidimensional Rasch model, item difficulty parameter values used in the simulation for the 43-items and 96-items conditions are the same as the ones used in the 6-tetlets and 12-testlets conditions, respectively

Appendix D: Expected Fisher Information Computation

For a testlet k consists of j items, the expected Fisher information is

$$I_k(\theta) = E_{\boldsymbol{y}_k}\left[-\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2}\right] = -\sum_{\boldsymbol{y}_k}\left(\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2} p(\boldsymbol{y}_k|\theta)\right),$$

where $p\left(\boldsymbol{y}_k|\hat{\theta}\right)$ is the marginal probability of score pattern $\boldsymbol{y}_k$ after marginalizing out the nuisance dimension, and $l(\theta|\boldsymbol{y}_k)$ is the log-likelihood of $\boldsymbol{y}_k$. The right-hand side of the above equation can be written as

$$-\sum_{\boldsymbol{y}_k}\left(\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2} p(\boldsymbol{y}_k|\theta)\right) = -\sum_{r_k=0}^{n_k}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2} p(\boldsymbol{y}_k|\theta),$$

where n_k is the number of items in the testlet, and $\mathbb{y}_{r_k}$ is the set of score patterns that leads to a sum score of r_k. Using the property that the sufficient statistic for θ is the raw score, we have $\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2} = \frac{d^2 l(\theta|r_k)}{d\theta^2}$ when $\boldsymbol{y}_k \in \mathbb{y}_{r_k}$. Hence

$$\begin{aligned}-\sum_{r_k=0}^{n_k}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\frac{d^2 l(\theta|\boldsymbol{y}_k)}{d\theta^2} p(\boldsymbol{y}_k|\theta) &= -\sum_{r_k=0}^{n_k}\left[\frac{d^2 l(\theta|r_k)}{d\theta^2}\sum_{\boldsymbol{y}_k\in\mathbb{y}_{r_k}}\left(p(\boldsymbol{y}_k|\theta)\right)\right]\\ &= -\sum_{r_k=0}^{n_k}\left[\frac{d^2 l(\theta|r_k)}{d\theta^2} p(r_k|\theta)\right].\end{aligned}$$

The computation of $p(r_k|\theta)$ is shown in the main body using the Lord-Wingersky algorithm. For $\frac{d^2 l(\theta|r_k)}{d\theta^2}$,

$$\begin{aligned}\frac{d^2 l(\theta|r_k)}{d\theta^2} = &\frac{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right)\left(r_k - \sum_{j=1}^{n_k} p_{jk}\right)^2 g\left(u_k|0,\sigma_{u_k}^2\right)du_k}{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right) g\left(u_k|0,\sigma_{u_k}^2\right)du_k} \quad \text{A}\\ &-\frac{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right)\left(\sum_{j=1}^{n_k} p_{jk}q_{jk}\right) g\left(u_k|0,\sigma_{u_k}^2\right)du_k}{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right) g\left(u_k|0,\sigma_{u_k}^2\right)du_k} \quad \text{B}\\ &-\left(\frac{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right)\left(r_k - \sum_{j=1}^{n_k} p_{jk}\right) g\left(u_k|0,\sigma_{u_k}^2\right)du_k}{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k}\log(q_{jk})\right) g\left(u_k|0,\sigma_{u_k}^2\right)du_k}\right)^2 \quad \text{C}\end{aligned}$$

Define $p(u_k|\theta,r_k) = \frac{\mathrm{Exp}\left(r_k u_k+\sum_{j=1}^{n_k}\log(q_{jk})\right) f\left(u_k|0,\sigma_{u_k}^2\right)}{\int \mathrm{Exp}\left(r_k u_k+\sum_{j=1}^{n_k}\log(q_{jk})\right) f\left(u_k|0,\sigma_{u_k}^2\right)du_k}$, then term A above is (omitting cluster index k)

$$\sum_{r=0}^{n}\left\{\int p(u|\theta,r)\left(r-\sum_{j=1}^{n} p_j\right)^2 du\right\} p(r) = \int\left\{\sum_{r=0}^{n} p(u|\theta,r)\, p(r)\left(r-\sum_{j=1}^{n} p_j\right)^2\right\} du$$

$$
\begin{aligned}
&= \int \left\{ \sum_{r=0}^{n} p\left(r, u|\theta\right) \left(r - \sum_{j=1}^{n} p_j \right)^2 \right\} du \\
&= \int \left\{ \sum_{r=0}^{n} p\left(r \,|\, \theta, u\right) p\left(u\right) \left(r - \sum_{j=1}^{n} p_j \right)^2 \right\} du \\
&= \int p\left(u\right) \left\{ \sum\nolimits_{r=0}^{n} p\left(r \,|\, \theta, u\right) \left(r - \sum\nolimits_{j=1}^{n} p_j \right)^2 \right\} du \\
&= \int p\left(u\right) E_{r|u} \left(r - \sum\nolimits_{j=1}^{n} p_j \right)^2 du.
\end{aligned}
$$

Since $E_{r|u}\left(r - \sum_{j=1}^{n} p_j\right)^2 = \sum_{j=1}^{n} p_j q_j$, term A becomes

$$
\int p\left(u\right) \left(\sum_{j=1}^{n} p_j q_j \right) du \ .
$$

Similarly, term B is,

$$
\begin{aligned}
&- \sum_{r=0}^{n} \left\{ \int p\left(u \,|\, \theta, r\right) \left(\sum_{j=1}^{n} p_j q_j \right) du \right\} p\left(r\right) \\
&= - \int \left\{ \sum_{r=0}^{n} p\left(u \,|\, \theta, r\right) p\left(r\right) \left(\sum_{j=1}^{n} p_j q_j \right) \right\} du \\
&= - \int \left\{ \sum_{r=0}^{n} p\left(r, u|\theta\right) \sum_{j=1}^{n} p_j q_j \right\} du = - \int \left\{ \sum_{r=0}^{n} p\left(r|\theta, u\right) p\left(u\right) \sum_{j=1}^{n} p_j q_j \right\} du \\
&= - \int p\left(u\right) \left(\sum_{j=1}^{n} p_j q_j \right) \left(\sum\nolimits_{r=0}^{n} p\left(r|\theta, u\right) \right) du = - \int p\left(u\right) \left(\sum_{j=1}^{n} p_j q_j \right) du.
\end{aligned}
$$

Therefore, term A and term B canceled out, and $\frac{d^2 l(\theta|r_k)}{d\theta^2}$ is simply provided by term C. Hence,

$$
I_k\left(\theta\right) = \sum_{r_k=0}^{n_k} \left[\left(\frac{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k} \log\left(q_{jk}\right) \right) \left(r_k - \sum_{j=1}^{n_k} p_{jk} \right) g\left(u_k \,|\, 0,\, \sigma_{u_k}^2 \right) du_k}{\int \mathrm{Exp}\left(r_k u_k + \sum_{j=1}^{n_k} \log\left(q_{jk}\right) \right) g\left(u_k \,|\, 0,\, \sigma_{u_k}^2 \right) du_k} \right)^2 p\left(r_k|\theta\right) \right] .
$$

It follows that the test information at $\theta = \hat{\theta}$ for a test with K testlets is

$$
I\left(\hat{\theta}\right) = \sum\nolimits_{k=1}^{K} I_k\left(\hat{\theta}\right).
$$

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.
Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika, 62*(2), 191–199.
Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168.
Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika, 80*(2), 535–559.
Chen, H. (2013). Testlet Effects on Standardized Log-likelihood Person Fit Index to Detect Aberrant Responses for the IRT Testlet Model (Doctoral dissertation, University of Missouri–Columbia).
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*(2), 159–177.
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.
Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*(2), 159–180.
Gorney, K., Sinharay, S., Eckerly, C. (2024). Efficient corrections for standardized person-fit statistics. Psychometrika, 1–23.
Hong, M., Lin, L., & Cheng, Y. (2021). Asymptotically corrected person fit statistics for multidimensional constructs with simple structure and mixed item types. *Psychometrika, 86*(2), 464–488.
Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298.
Liou, M., & Chang, C. H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57*(2), 169–181.
Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*(4), 453–461.
Magis, D., Raîche, G., & Béland, 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*(1), 57–81.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.
Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75–106.
Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*(2), 121–129.
New Hampshire Department of Education (2019). New hampshire statewide assessment system 2018-2019 annual technical report volume 1. https://www.education.nh.gov/sites/g/files/ehbemt326/files/inline-documents/sonh/nhsas-v1-tech-report-2018-19.pdf
Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*(3), 213–229.
Rijmen, F., Turhan, A., Jiang, T. (2018). An item response theory model for next generation of science standards assessments. National Council of Measurement in Education Annual Conference, New York, NY.
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.
Seo, D. G., & Weiss, D. J. (2013). lz Person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73*(6), 994–1016.
Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics, 40*(4), 343–365.
Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81*(4), 992–1013.
Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.
Snijders, T. A. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.
van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*(4), 327–345.
von Davier, M., & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*(2), 213–228.
Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*(5), 741–758.
Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.
Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203–220.
Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

Xia, Y., & Zheng, Y. (2018). Asymptotically normally distributed person fit indices for detecting spuriously high scores on difficult items. *Applied Psychological Measurement, 42*(5), 343–358.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Manuscript Received: 3 JUN 2022
Final Version Received: 2 MAY 2024
Published Online Date: 17 AUG 2024