Extension of caution indices to mixed-format tests

Sandip Sinharay, Educational Testing Service

An Updated Version of this document appeared on 01/09/2018 in the British Journal of Mathematical and Statistical Psychology. The website for the article is https://onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12124

The citation for the article is: Sinharay, S. (2018). Extension of caution indices to mixed-format tests. British Journal of Mathematical and Statistical Psychology. doi:10.1111/bmsp.12124

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

Extension of Caution Indices to Mixed-format Tests

Sandip Sinharay, Educational Testing Service

October 11, 2017

Extension of Caution Indices to Mixed-format Tests

Abstract

Tatsuoka (1984) suggested several extended caution indices and their standardized versions that have been used as person-fit statistics by researchers such as Drasgow, Levine, and McLaughlin (1987), Glas and Meijer (2003), and Molenaar and Hoijtink (1990). However, these indices are only defined for tests with dichotomous items. This paper extends two of the popular standardized extended caution indices (Tatsuoka, 1984) for use with polytomous items and mixed-format tests. Two additional new person-fit statistics are obtained by applying the asymptotic standardization of person-fit statistics for mixed-format tests (Sinharay, 2016c). Detailed simulations are then performed to compute the Type I error rate and power of the four new person-fit statistics. Two real data illustrations follow. The new person-fit statistics appear to be satisfactory tools for assessing person fit for polytomous items and mixed-format tests.

Key words: caution index, person fit, ζ_1 statistic, ζ_2 statistic.

Acknowledgements

The author would like to thank the editor, Matthias von Davier, the associate editor, Bernard Veldkamp, and the two anonymous reviewers for several helpful comments that led to a significant improvement of the paper. The author would also like to thank James Wollack for generously sharing two data sets one of which was used in the research that led to this paper. The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

Person-fit assessment (PFA) is concerned with uncovering atypical test performance as reflected in the pattern of scores on individual items in a test (Meijer & Sijtsma, 2001). In a report for the Council of Chief State School Officers, Olson and Fremer (2013) recommended the use of person-fit statistics (PFSs), in addition to other methods, to detect irregularities in answering behavior.

Several PFSs have been proposed in the context of tests with dichotomous items—see comprehensive reviews of them by, for example, Karabatsos (2003) and Meijer and Sijtsma (2001). There exist fewer PFSs (e.g., Drasgow, Levine, & Williams, 1985; Emons, 2008; Glas & Dagohoy, 2007; van Krimpen-Stoop & Meijer, 2002; von Davier & Molenaar, 2003; Wright & Masters, 1982) for tests with polytomous items.

There is a severe lack of research on PFA for mixed-format tests (MFTs), which are tests that include both dichotomous and polytomous items, Finkelman and Kim (2007), Sinharay (2016c), Sinharay (2015), and Tendeiro (2017) being some exceptions. Polytomous items and MFTs promise to become more common because of an increasing emphasis on constructed-response items in the common-core assessments (e.g., Darling-Hammond & Adamson, 2010, p. 1). Constructed-response items constitute an integral part of the assessment design of both of the federally-funded assessment consortia—the Smarter Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Career (e.g., Lissitz, Hou, & Slater, 2012). Therefore, research on PFA for polytomous items and MFTs promise to be useful to educational and psychological measurement.

Tatsuoka (1984) suggested several extended caution indices (ECI) and their standardized versions (ECI_z) that can be used as PFSs. The standardized versions of the second ECI (denoted as $ECI2_z$ or ζ_1) and the fourth ECI ($ECI4_z$ or ζ_2) are arguably the most popular among the caution indices and their extensions, which is evident from the fact that several researchers (e.g., Drasgow et al., 1987; Glas & Meijer, 2003; Karabatsos, 2003; Molenaar & Hoijtink, 1990) found ζ_1 and ζ_2 to be useful in detecting person misfit. However, ζ_1 and ζ_2 can be applied only to tests with dichotomous items. Sinharay (2015) suggested an extension of ζ_2 to MFTs, but the extended PFS had low power.

This paper suggests four new PFSs including two extensions each of ζ_1 and ζ_2 for use with polytomous items and MFTs. Two of the new PFSs are based on the asymptotic standardization/correction of PFSs for MFTs recently suggested by Sinharay (2016c).

The Literature Review section includes reviews of the ζ_1 and ζ_2 statistics for dichotomous items (Tatsuoka, 1984), the extension of the ζ_2 statistic suggested by Sinharay (2015), and the asymptotic standardization/correction of PFSs for MFTs (Sinharay, 2016c). The Methods section includes the descriptions of the new PFSs. The Type I error rate and power of the new PFSs are examined for several simulated data sets in the Simulations section. In the Application section, the suggested PFSs are computed for two real data sets. Conclusions and recommendations are provided in the last section.

The new PFSs are based on item response theory (IRT). Non-parametric PFSs (for example, Emons, 2008; Sijtsma, 1998; Tendeiro & Meijer, 2014) are not considered in this paper.

Literature Review

Review of the ζ_1 and ζ_2 Statistics for Dichotomous Items

Consider a test comprising J dichotomous items that was administered to n examinees. Let us denote the true ability of examinee i as θ_i. Let y_{ij} be the score (that is 0 or 1) and $P_j(\theta_i)$ be the probability that y_{ij} is equal to 1 for examinee i on item j. For example, for the three-parameter logistic model (3PLM),

$$P_j(\theta_i) = P(y_{ij} = 1) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \tag{1}$$

where a_j, b_j, and c_j respectively are the (known) slope, difficulty, and guessing parameters of item j.

Let us define

$$G_j = \frac{1}{n}\sum_{i=1}^{n} P_j(\theta_i), j = 1, 2, \cdots, J, \text{ and } G = \frac{1}{J}\sum_{j=1}^{J} G_j. \tag{2}$$

Let us further define

$$\sigma_j^2(\theta_i) = \text{Var}(y_{ij}) = P_j(\theta_i)[1 - P_j(\theta_i)] \text{ and } \bar{P}(\theta_i) = \frac{1}{J}\sum_{j=1}^{J} P_j(\theta_i).$$

For the i-th examinee, Tatsuoka (1984) defined the second standardized ECI, denoted as $ECI2_z$ or ζ_1, as

$$\zeta_1 = \frac{\sum_{j=1}^{J}(P_j(\theta_i) - y_{ij})(G_j - G)}{\sqrt{\sum_{j=1}^{J}\sigma_j^2(\theta_i)(G_j - G)^2}}, \tag{3}$$

and the fourth standardized ECI, denoted as $ECI4_z$ or ζ_2, as

$$\zeta_2 = \frac{\sum_{j=1}^{J}(P_j(\theta_i) - y_{ij})\left[P_j(\theta_i) - \bar{P}(\theta_i)\right]}{\sqrt{\sum_{j=1}^{J}\sigma_j^2(\theta_i)\left[P_j(\theta_i) - \bar{P}(\theta_i)\right]^2}}. \tag{4}$$

Note that the denominator in each of Equations 3 and 4 is the standard deviation (SD) of the numerator. Note also that the $ECI6_z$ statistic of Tatsuoka (1984) is identical to $ECI4_z$—so, ζ_2 denotes both $ECI4_z$ and $ECI6_z$. To compute ζ_1 and ζ_2 for a data set, θ_i has to be replaced by an estimate $\hat{\theta}_i$.

Tatsuoka (1984) assumed the asymptotic null distribution of ζ_1 or ζ_2, with θ_i replaced by $\hat{\theta}_i$, to be standard normal. In addition, either of ζ_1 and ζ_2 becomes larger as an examinee answers more difficult items correctly [1] or answers more easy items incorrectly, which usually happens when person misfit occurs. Therefore, a large value such as a value larger than 1.645 at 5% level of either of ζ_1 or ζ_2 indicates person misfit.

Both ζ_1 and ζ_2 have been used as PFSs by several researchers such as Drasgow et al. (1987), Glas and Meijer (2003), Karabatsos (2003), Li and Olenik (1997), and Molenaar and Hoijtink (1990). The ζ_2 statistic along with the (Bayesian) posterior predictive model checking method performed the best overall among eight PFSs in Glas and Meijer (2003). Drasgow et al. (1987) found ζ_1 and ζ_2 to have satisfactory power in a comparison of several PFSs. Sinharay (2016b) used the results of Snijders (2001) to suggest asymptotically standardized versions of the ζ_1 and ζ_2 for dichotomous items and found those standardized versions to have satisfactory Type I error rate and power. However, ζ_1 and ζ_2 are defined only for dichotomous items. Therefore, keeping in mind the increasing importance of polytomous items and MFTs, extensions of ζ_1 and ζ_2 to polytomous items and to MFTs may be helpful to researchers interested in PFA.

[1] For a difficult item, $G_j - G < 0$ and $P_j(\theta_i) - \bar{P}(\theta_i) < 0$. A correct answer on the item means $P_j(\theta_i) - y_{ij} < 0$ so that both $(P_j(\theta_i) - y_{ij})(G_j - G)$ and $(P_j(\theta_i) - y_{ij})\left[P_j(\theta_i) - \bar{P}(\theta_i)\right]$ are positive.

Notation for Mixed-format Tests

Consider a test with J items each of which can be dichotomous or polytomous. Consider the i-th examinee whose true ability is θ_i. The examinee's score on item j, y_{ij}, can be an integer between 0 and m_j. Let us denote, for $k = 0, 1, \ldots, m_j$,

$$d_k(y_{ij}) = \begin{cases} 1 \text{ if } y_{ij} = k \\ 0 \text{ otherwise} \cdot \end{cases} \tag{5}$$

$$\text{Let } P_{jk}(\theta_i) = P(y_{ij} = k) = P(d_k(y_{ij}) = 1) = E(d_k(y_{ij})) \tag{6}$$

denote the probability that the score of examinee i on item j is equal to k. Equations 5 and 6 subsume items modeled by any common IRT model for appropriate choices of $P_{jk}(\theta_i)$. For example, if item j is polytomous and if the generalized partial credit model (GPCM; Muraki, 1992) is used for the item, then Equations 5 and 6 subsume the item (and the GPCM) for

$$P_{jk}(\theta_i) = \frac{\exp[\sum_{h=0}^{k} a_j(\theta_i - b_{jh})]}{\sum_{c=0}^{m_j} \exp[\sum_{h=0}^{c} a_j(\theta - b_{jh})]},$$

where a_j's are the slope parameters and b_{jh}'s are the location parameters. If item j is dichotomous, then m_j becomes equal to 1, and, as a result, y_{ij} can be 0 or 1; then, for example, if the 3PLM is used for the item, then Equations 5 and 6 subsume the item (and the 3PLM) for

$$d_0(y_{ij}) = 1 - y_{ij}, d_1(y_{ij}) = y_{ij}, P_{j0}(\theta_i) = 1 - P_j(\theta_i), \text{ and } P_{j1}(\theta) = P_j(\theta_i), \tag{7}$$

where $P_j(\theta_i)$ was defined in Equation 1.

Note that MFTs include as special cases tests with only dichotomous items, tests with only polytomous items with the same number of response categories, and tests with only polytomous items with varying number of response categories (because m_j is allowed to vary over j).

Review of the Extension of ζ_2 to MFTs (Sinharay, 2015)

Sinharay (2015) suggested the following extended version of ζ_2 for use with MFTs:

$$\tilde{\zeta}_2 = \frac{\sum_{j=1}^{J} \left[\frac{E(y_{ij}|\theta_i)}{m_j} - \frac{y_{ij}}{m_j}\right] \left[\frac{E(y_{ij}|\theta_i)}{m_j} - U(\theta_i)\right]}{\sqrt{\sum_{j=1}^{J} \frac{V(y_{ij}|\theta_i)}{m_j^2} \left[\frac{E(y_{ij}|\theta_i)}{m_j} - U(\theta_i)\right]^2}}, \tag{8}$$

where $U(\theta_i) = \frac{1}{J}\sum_{j=1}^{J} \frac{E(y_{ij}|\theta_i)}{m_j}$. The statistic $\tilde{\zeta}_2$ was found to have considerably smaller power (often by a margin of 0.10 under some conditions) in the simulations in this paper compared to the extensions suggested later—so $\tilde{\zeta}_2$ is not considered in the remaining of this paper.

Review of the Asymptotic Standardization of PFSs for MFTs

Sinharay (2016c) considered, for MFTs, only those PFSs that are of the form

$$\frac{T(\theta_i)}{\sqrt{\text{Var}(T(\theta_i))}} \tag{9}$$

$$\text{for } T(\theta_i) = \sum_{j=1}^{J} \sum_{k=0}^{m_j} [d_k(y_{ij}) - P_{jk}(\theta_i)] w_{jk}(\theta_i), \tag{10}$$

where $w_{jk}(\theta_i)$ is a real-valued weight function. For example, the l_z statistic for polytomous items (Drasgow et al., 1985) is of the form given in Equation 9 for $w_{jk}(\theta_i) = \log P_{jk}(\theta_i)$.

The variance $\text{Var}(T(\theta_i))$ is equal to $\sum_{j=1}^{J} \mathbf{w}_j'(\theta_i)\mathbf{D_j}(\theta_i)\mathbf{w}_j(\theta_i)$, where

$$\mathbf{w}_j(\theta_i) = (w_{j0}(\theta_i), w_{j1}(\theta_i), \ldots, w_{jm_j}(\theta_i))'$$

$$\text{and } \mathbf{D_j}(\theta_i) = \text{the covariance matrix of } (d_0(y_{ij}), d_1(y_{ij}), \ldots, d_{m_j}(y_{ij}))'.$$

The covariance matrix $\mathbf{D_j}(\theta_i)$ is given by

$$\mathbf{D_j}(\theta_i) = \begin{pmatrix} P_{j0}(\theta_i)(1 - P_{j0}(\theta_i)) & -P_{j0}(\theta_i)P_{j1}(\theta_i) & \ldots & -P_{j0}(\theta_i)P_{jm_j}(\theta_i) \\ -P_{j1}(\theta_i)P_{j0}(\theta_i) & P_{j1}(\theta_i)(1 - P_{j1}(\theta_i)) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ -P_{jm_j}(\theta_i)P_{j0}(\theta_i) & -P_{jm_j}(\theta_i)P_{j1}(\theta_i) & \ldots & P_{jm_j}(\theta_i)(1 - P_{jm_j}(\theta_i)) \end{pmatrix}.$$

Sinharay (2016c) showed that the asymptotic null distribution of the PFS in Equation 9, with θ_i replaced by an estimate $\hat{\theta}_i$, is not $\mathcal{N}(0,1)$, but that of

$$\frac{T(\hat{\theta}_i) + c'_n(\hat{\theta}_i)s_0(\hat{\theta}_i)}{\sqrt{\widetilde{\text{Var}}(T(\hat{\theta}_i))}} \tag{11}$$

is $\mathcal{N}(0,1)$ under three regularity conditions, where

$$\widetilde{\text{Var}}(T(\hat{\theta}_i)) = \sum_{j=1}^{J} \mathbf{v}'_j(\hat{\theta}_i)\mathbf{D_j}(\hat{\theta}_i)\mathbf{v}_j(\hat{\theta}_i), \tag{12}$$

$$\text{for } \mathbf{v}_j(\theta_i) = (\tilde{w}_{j0}(\theta_i), \tilde{w}_{j1}(\theta_i), \ldots, \tilde{w}_{jm_j}(\theta_i))',$$

that is, $\widetilde{\text{Var}}(T(\hat{\theta}_i))$ is obtained by replacing $w_{jk}(\hat{\theta}_i)$ in the expression of $\text{Var}(T(\hat{\theta}_i))$ by $\tilde{w}_{jk}(\hat{\theta}_i)$, where

$$\tilde{w}_{jk}(\hat{\theta}_i) = w_{jk}(\hat{\theta}_i) - c'_n(\hat{\theta}_i)s_{jk}(\hat{\theta}_i), \tag{13}$$

$$c'_n(\hat{\theta}_i) = \frac{\sum_{j=1}^{J}\sum_{k=0}^{m_j} P'_{jk}(\hat{\theta}_i)w_{jk}(\hat{\theta})}{\sum_{j=1}^{J}\sum_{k=0}^{m_j} P'_{jk}(\hat{\theta})s_{jk}(\hat{\theta}_i)}, \tag{14}$$

and $\hat{\theta}_i$ satisfies

$$s_0(\hat{\theta}_i) + \sum_{j=1}^{J}\sum_{k=0}^{m_j}[d_k(y_{ij}) - P_{jk}(\hat{\theta}_i)]s_{jk}(\hat{\theta}_i) = 0 \tag{15}$$

for some functions $s_0(\hat{\theta}_i)$ and $s_{jk}(\hat{\theta}_i)$, where

$$P'_{jk}(\hat{\theta}_i) = \text{the first derivative of } P_{jk}(\hat{\theta}_i) \text{ with respect to } \hat{\theta}_i. \tag{16}$$

Sinharay (2016c) suggested the l_z^* statistic for MFTs that is an extension of the l_z^* statistic for dichotomous items (Snijders, 2001) and is of the form given by Equation 11 for $w_{jk}(\theta_i) = \log P_{jk}(\theta_i)$. Expressions of $P'_{jk}(\hat{\theta}_i)$ for the common IRT models such as the 3PLM and GPCM can be found in, for example, Tao, Shi, and Chang (2012). The condition provided in Equation 15 is satisfied by the maximum likelihood estimate (MLE), weighted maximum likelihood estimate (WLE; Warm, 1989), and modal a posteriori (MAP) estimate of ability, and, for all these estimates,

$$s_{jk}(\hat{\theta}_i) = \frac{P'_{jk}(\hat{\theta}_i)}{P_{jk}(\hat{\theta}_i)}. \tag{17}$$

The quantity $s_0(\hat{\theta}_i)$ can be computed as

$$s_0(\hat{\theta}_i) = \begin{cases} 0 \text{ if } \hat{\theta}_i = \text{MLE}, \\ \frac{d \log f(\hat{\theta}_i)}{d\hat{\theta}_i} \text{ if } \hat{\theta}_i = \text{MAP}, \\ \frac{\mathcal{J}(\hat{\theta}_i)}{2I(\hat{\theta}_i)} \text{ if } \hat{\theta}_i = \text{WLE}, \end{cases} \tag{18}$$

where $f(\theta_i)$ is the prior distribution on θ_i, $I(\theta_i)$ is the information on θ_i, and $\mathcal{J}(\theta_i) = \sum_j \sum_k \frac{P'_{jk}(\theta_i) P''_{jk}(\theta_i)}{P'_{jk}(\theta_i)}$, where $P''_{jk}(\theta_i)$ is the second derivative of $P_{jk}(\theta_i)$. The results of Sinharay (2016c) are extensions of similar results suggested for dichotomous items by Snijders (2001), who applied his results to derive the standardized l_z^* statistic from the unstandardized l_z statistic (Drasgow et al., 1985) for dichotomous items. The three regularity conditions of Sinharay (2016c) are satisfied by all common IRT models for MFTs including combinations of the 1-, 2-, and 3-parameter logistic and probit models, the GPCM, the partial credit model (Masters, 1982), and the graded response model (Samejima, 1969). Tendeiro (2017) applied the extension of Sinharay (2016c) to suggest a PFS $l^*_{z(p)}$ that can be applied to unfolding models.

Methods

Extensions of ζ_1 and ζ_2 to Mixed-format Tests

Let us define, for $k = 0, 1, \ldots, m_j$,

$$G_{jk} = \frac{1}{n} \sum_{i=1}^{n} P_{jk}(\theta_i) \text{ and } G_k = \frac{1}{J_k} \sum_{j \in S_k} G_{jk},$$

where S_k is the set of items that have a score category of k and J_k is the size of S_k. Let us define

$$\bar{P}_k(\theta_i) = \frac{1}{J_k} \sum_{j \in S_k} P_{jk}(\theta_i) \cdot$$

Usually, J_k would be smaller than J. If all the items have the same number of score categories, then $J_k = J$ and S_k is the set of all items.

Then, for examinee i, an extended version of ζ_1 for use with MFTs can be obtained as

$$\zeta_1 = \frac{\sum_{j=1}^{J} \sum_{k=0}^{m_j} [P_{jk}(\theta_i) - d_k(y_{ij})][G_{jk} - G_k]}{\sqrt{\sum_{j=1}^{J} \mathbf{f}'_{\mathbf{j}}(\theta_i) \mathbf{D}_{\mathbf{j}}(\theta_i) \mathbf{f}_{\mathbf{j}}(\theta_i)}}, \tag{19}$$

$$\text{where } \mathbf{f_j}(\theta_i) = (G_{j0} - G_0, G_{j1} - G_1, \cdots, G_{jm_j} - G_{m_j})',$$

and an extended version of ζ_2 for use with MFTs can be obtained as

$$\zeta_2 = \frac{\sum_{j=1}^{J} \sum_{k=0}^{m_j} [P_{jk}(\theta_i) - d_k(y_{ij})][P_{jk}(\theta_i) - \bar{P}_k(\theta_i)]}{\sqrt{\sum_{i=1}^{J} \mathbf{g}'_{\mathbf{j}}(\theta_i)\mathbf{D_j}(\theta_i)\mathbf{g_j}(\theta_i)}}, \tag{20}$$

$$\text{where } \mathbf{g_j}(\theta_i) = (P_{j0}(\theta_i) - \bar{P}_0(\theta_i), P_{j1}(\theta_i) - \bar{P}_1(\theta_i), \cdots, P_{jm_j}(\theta_i) - \bar{P}_{m_j}(\theta_i))'.$$

A comparison of Equations 3 and 19 (or of Equations 4 and 20) shows that

- the probability of a correct answer in Equation 3 (or 4), $P_j(\theta_i)$, is replaced in Equation 19 (or 20) by $P_{jk}(\theta_i)$, the probability of a score equal to a specific score category.
- the binary item score y_{ij} in Equation 3 (or 4) is replaced in Equation 19 (or 20) by the binary category score indicator $d_k(y_{ij})$.
- G in Equation 3 is replaced by G_k in Equation 19 and $\bar{P}(\theta_i)$ in Equation 4 is replaced by $\bar{P}_k(\theta_i)$ in Equation 20.
- As in Equation 3 (or 4), the denominator of Equation 19 (or 20) is the SD of the numerator.

Thus, the extended versions of ζ_1 and ζ_2 for use with MFTs capture person misfit in the same manner in which ζ_1 and ζ_2 capture misfit for dichotomous items, but involve appropriate adjustments for polytomous items. The value of either of the extended version of ζ_1 and ζ_2 is expected to be large and positive in the presence of person misfit.

A comparison of Equations 8 and 20 makes it clear that the former does not involve a summation over each possible score category of the items whereas the latter does. Thus, the statistic provided in Equation 8 involves the loss of some information, which may be the reason of its smaller power compared to that provided in Equation 20.

To compute the above extended versions of ζ_1 and ζ_2 for a data set, θ_i has to be replaced by an estimate $\hat{\theta}_i$. If all the items on the test are dichotomous, then the right-hand side of Equation 19 becomes equal to that of Equation 3 and the right-hand side of Equation 20

becomes equal to that of Equation 4; a proof, which involves some algebra, is omitted, but the proof is similar to the proof of Theorem 1 in Sinharay (2016c) and can be obtained upon request from the author.

Asymptotic Standardization of the Suggested Extensions

The extended versions of ζ_1 (Equation 19) and ζ_2 (Equation 20) are both special cases of the PFS that was considered in Sinharay (2016c) and can be expressed using Equation 9 for

$$w_{jk}(\theta_i) = -(G_{jk} - G_k) \text{ and } w_{jk}(\theta_i) = -(P_{jk}(\theta_i) - \bar{P}_k(\theta_i)), \tag{21}$$

respectively. Thus, when θ_i is replaced by its MLE, WLE, or MAP, denoted as $\hat{\theta}_i$, the asymptotic standardization/correction suggested by Sinharay (2016c) can be applied to either of ζ_1 and ζ_2 to obtain the corresponding asymptotic standardized/corrected versions. Let us denote these standardized/corrected versions as ζ_1^* and ζ_2^*, respectively, that is,

$$\zeta_1^* = \frac{T(\hat{\theta}_i) + c'_n(\hat{\theta}_i)s_0(\hat{\theta}_i)}{\sqrt{\widetilde{\text{Var}}(T(\hat{\theta}_i))}} \text{ where } w_{jk}(\hat{\theta}_i) = -(G_{jk} - G_k), \tag{22}$$

$$\text{and } \zeta_2^* = \frac{T(\hat{\theta}_i) + c'_n(\hat{\theta}_i)s_0(\hat{\theta}_i)}{\sqrt{\widetilde{\text{Var}}(T(\hat{\theta}_i))}} \text{ where } w_{jk}(\hat{\theta}_i) = -(P_{jk}(\theta_i) - \bar{P}_k(\theta_i)), \tag{23}$$

and $T(\theta_i)$, $\widetilde{\text{Var}}(T(\hat{\theta}_i))$, $c'_n(\hat{\theta}_i)$, and $s_0(\hat{\theta}_i)$ are defined in Equations 10, 12, 14, and 18, respectively.

Asymptotic Null Distributions of ζ_1^* and ζ_2^*

Sinharay (2016c) proved in the context of MFTs that if a PFS is of the form given by Equation 9 for some $w_{jk}(\theta_i)$, then the asymptotic null distribution of its standardized version provided in Equation 11 is $\mathcal{N}(0, 1)$ under three regularity conditions. Because both ζ_1 and ζ_2 for MFTs are of the form given by Equation 9 for $w_{jk}(\theta_i)$'s given by Equation 21, the proof of Sinharay (2016c) implies that ζ_1 and ζ_2 do not have a standard normal asymptotic null distribution, but that their standardized versions ζ_1^* and ζ_2^* provided in

Equations 22 and 23 have a standard normal asymptotic null distribution under three regularity conditions. The regularity conditions are mild and satisfied by the IRT models commonly used for MFTs.

A test with only polytomous items with the same number of response categories for all items or a test with only polytomous items with a varying number of response categories is a special case of a MFT; thus, the suggested PFSs ζ_1^* and ζ_2^* have a standard normal asymptotic null distribution for such tests as well. If a test includes only dichotomous items, then the suggested ζ_1^* and ζ_2^* statistics becomes identical to the correponding asymptotically standardized versions (for dichotomous items) suggested in Sinharay (2016b) and has a standard normal asymptotic null distribution. Sinharay (2016c) showed that the PFSs obtained by removing the term $c'_n(\hat{\theta}_i)s_0(\hat{\theta}_i)$ from the numerator of Equation 11 also have a standard normal asymptotic null distribution; however, the term $c'_n(\hat{\theta}_i)s_0(\hat{\theta}_i)$ is included in the remaining of this paper.

Computations

Given the data, an IRT model, and the estimated item parameters, the computation of ζ_1^* and ζ_2^* for an examinee involves the computation of the quantities/expressions provided in the left column Table 1, in the same order as that of the rows of the table, using the formulas provided in the right column of the table.

A Simulation Study

The Type I error rate and power of ζ_1, ζ_2, ζ_1^* and ζ_2^* were examined for a variety of simulated MFTs. The simulation study also included the l_z^* statistic (Sinharay, 2016c) that is another asymptotically corrected PFS (like ζ_1^* and ζ_2^*) for use with MFTs.

Design of the Simulation

The simulation study involved three levels of test length (12 items, 30 items, and 60 items) that represented short, moderate, and long tests. Each generated data set involved two sets of items, a set of dichotomous items and a set of polytomous items, which resulted

Table 1. The quantities that need to be computed in order to compute ζ_1^* and ζ_2^*.

Compute	Using
$\hat{\theta}_i$ (MLE, WLE, or MAP)	The scores, the item parameters and a maximization algorithm such as the Newton-Raphson algorithm
$d_k(y_{ij})$	Equation 5
$P_{jk}(\hat{\theta}_i)$	Equation 6
$s_{jk}(\hat{\theta}_i)$	Equation 17
$s_0(\hat{\theta}_i)$	Equation 18
$w_{jk}(\hat{\theta}_i)$	Equation 21
$P'_{jk}(\hat{\theta}_i)$	Equation 16
$c'_n(\hat{\theta}_i)$	Equation 14
$\tilde{w}_{jk}(\hat{\theta}_i)$	Equation 13
$T(\hat{\theta}_i)$	Equation 10
$\widetilde{\mathrm{Var}}(T(\hat{\theta}_i))$	Equation 12
ζ_1^* and ζ_2^*	Equations 22 and 23

in the data set being like one arising from a MFT. The number of polytomous items was 4, 10, and 20, respectively (that is, one-third), for the three test lengths. The number of response categories for each polytomous item was fixed at three with possible scores being 0, 1, and 2. Scores on dichotomous and polytomous items were generated using the 3PLM and GPCM, respectively. The true slope parameters of all items were generated, as in Glas and Dagohoy (2007), from a log-normal distribution with respectively 0 and 0.25 as the mean and SD of the logarithm of the variable. The true difficulty and true guessing parameters for the dichotomous items were generated from a $\mathcal{N}(0,1)$ and a Uniform(0.05,0.3) distribution, respectively. The true location parameters of the polytomous items were generated from $\mathcal{N}(-1,0.5)$ and $\mathcal{N}(1,0.5)$ distributions, respectively, as in Chon, Lee, and Dunbar (2010).

To compute the Type I error of the PFSs, score patterns that fit the IRT (3PLM+GPCM) model were generated. To compute the power of the PFSs, score patterns that are "corrupted" and do not fit the IRT model were generated in several ways. The item parameters are assumed known; because of this assumption, the power does not depend on the number of examinees in a data set whose score patterns were corrupted—so the score patterns of all examinees were corrupted in each data set used to compute power.

As in other simulation studies on PFA (e.g., Glas & Meijer, 2003; Sinharay, 2016c; van Krimpen-Stoop & Meijer, 2002), corrupted score patterns reflected "lack of motivation" or "item disclosure/preknowledge". When "lack of motivation" was simulated, the score patterns of all examinees involved lack of motivation on $\frac{1}{3}$ or $\frac{1}{6}$ of all items. It was assumed, as in, for example, Glas and Meijer (2003), that the dichotomous items on which an examinee lacks motivation are the easiest among all the dichotomous items. The probability of a correct response to a dichotomous item on which an examinee lacks motivation was set to 0.2 as in Glas and Meijer (2003). For a polytomous item under "lack of motivation", 2.5 was subtracted from the examinee ability before generating a score on the item—it was found from a preliminary simulation that this reduction of 2.5 was somewhat equivalent on an average to setting the probability of a correct answer on a dichotomous item to 0.2. [2] When "item disclosure" was simulated, the score patterns of all examinees involved the assumption that $\frac{1}{3}$ or $\frac{1}{6}$ of all items were disclosed to the examinee. It was assumed, as in Glas and Meijer (2003), that the dichotomous items on which item disclosure occurs are the most difficult among all the dichotomous items. The score on a disclosed item (dichotomous or polytomous) was set equal to m_j, the highest possible score on the item.

For each simulation condition (where an example of a simulation condition is "12 items and lack of motivation on $\frac{1}{6}$ items"), 1,000 data sets with 1,000 examinees each were simulated; the true item parameters were simulated once for each of the 1,000 data sets. The true θ_i's of the examinees were uniformly drawn from one of the following 9 values: –2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, and 2.0. This design, which was also used by researchers such as Snijders (2001) and van Krimpen-Stoop and Meijer (1999), allows the computation of the Type I error rate and power at these θ-values accurately. The Type I error rates at levels of 1% and 5% for a PFS at a specific θ-value for a test length was computed as the proportion of score patterns fitting the IRT model under that test length

[2]Consider a dichotomous item with 1, 0, and 0.15 as the estimated slope, difficulty, and guessing parameters—these values are all close to the average of the generating item parameters—so the item can be considered as an average item. The probability of a correct answer on the item is 0.58 for ability 0 (average ability). To make the probability of a correct answer equal to 0.2, the ability has to be smaller than -2.6.

with that θ-value for which the PFS was statistically significant (under a standard normal null distribution assumption). Thus, the Type I error rate at any of the nine values of θ_i is computed from about 111,111 ($\approx 1000\times1000/9$) examinees for any given test length. The standard error is approximately 0.0003 when the Type I error is close to 0.01 and 0.0007 when the Type I error is close to 0.05; that is because, for example, when the Type I error is close to 0.01, the corresponding standard error is equal to $\sqrt{(0.01 \times 0.99)/111111} \approx 0.0003$. The power at 5% level for a PFS at a θ-value for a simulation condition was computed as the proportion of the corrupted score patterns under that simulation condition with that θ-value for which the PFS was statistically significant (under a standard normal null distribution assumption) at that level. Thus, the power at each ability was computed from about 111,111 ($\approx$1,000$\times$1,000/9) examinees in any simulation condition. The standard error for any value of power is always smaller than 0.0015 (the maximum occurring near power values of 0.5).

Computations

Fortran 90 programs written by the author were used for the computation of the estimates of ability and the PFSs.

For any simulation condition, the following steps were repeated 1,000 times:

1. Simulate a set of true item parameters; simulate 1,000 true ability values (representing 1,000 examinees) uniformly from the nine above-mentioned values.

2. Use the above true item and ability parameters and the IRT model (3PLM+GPCM) to simulate the item scores on a data set; simulate score patterns from the IRT model for simulation conditions to compute the Type I error rate and simulate corrupted score patterns for simulation conditions to compute the power.

3. Compute $\hat{\theta}_i$ of all the examinees from the data set using the true item parameters.

4. Compute ζ_1, ζ_2, ζ_1^*, ζ_2^*, and l_z^* for each examinee in the data set using $\hat{\theta}_i$ and the true item parameters. Compute the p-values corresponding to these PFSs under a standard normal null distribution assumption.

Results: Null Distributions of the PFSs

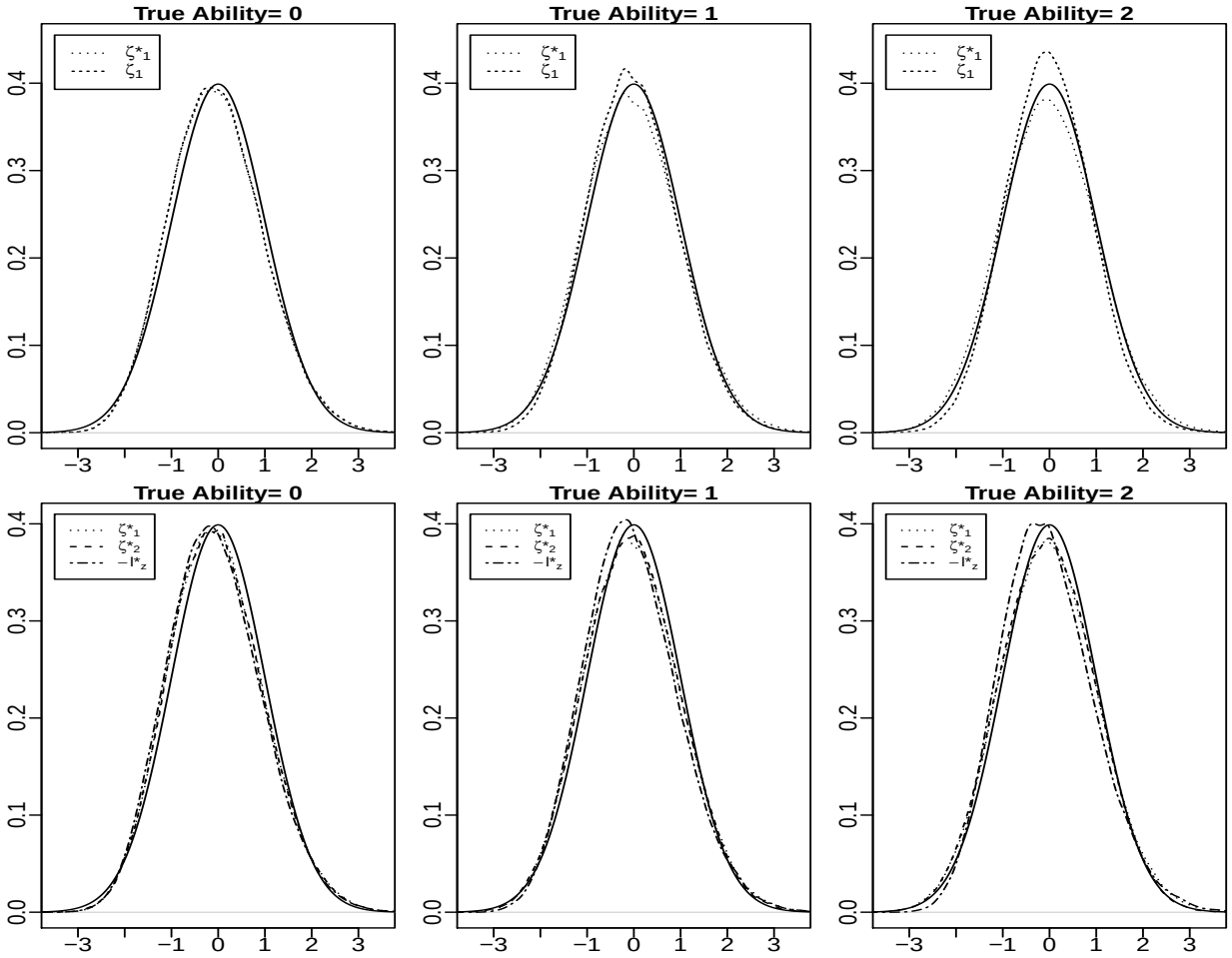

Figure 1. Null Distributions of the PFSs for three values of ability for the 60-item test.

The top three panels of Figure 1 show the distributions of the observed values of ζ_1^* (a line with dots of regular font) and ζ_1 (line with bold dots) computed from all the response patterns that fit the IRT model and were simulated using true abilities 0, 1, and 2 (mentioned in the titles of the panels) for 60-item tests. The distributions for true ability values of -1 and -2 are very similar to those for true ability values of 1 and 2, respectively, and are not shown here. The density of the standard normal distribution is also shown using a solid line. The top panels of the figure show that the distributions of ζ_1 and ζ_1^* are virtually indistinguishable for ability=0, but differ for non-zero values of the ability. The top panels of the figure also show that the right tail of the distribution of ζ_1 is lighter

than that of the standard normal distribution for values of the PFS between 1.0 and 2.0 (roughly)—this phenomenon is expected to manifest itself as ζ_1 being conservative for true ability of 1.0 or more at 5% level (note that 1.645, the 95th percentile of the standard normal distribution, lies between 1.0 and 2.0). In contrast, the right tail of the distribution of ζ_1^* follows that of the standard normal distribution more closely for values between 1.0 and 2.0—so the Type I error rates of ζ_1^* are expected to be close to the nominal level at 5% level. However, for values of the PFSs larger than 2 (that is close to the 2nd percentile of the standard normal distribution), the distribution of ζ_1^* has a heavier tail compared to the standard normal distribution—so the PFS is expected to have a Type I error rate slightly larger than the nominal level at levels around 1%.

The bottom three panels of Figure 1 show the distributions of the observed values of ζ_1^*, ζ_2^*, and $-l_z^*$ (the negative of $-l_z^*$ is plotted so that a large value of each PFS plotted in these panels indicates a person misfit) for true abilities of 0, 1, and 2. The standard normal distribution is also shown as a solid line. The three distributions are quite close in the bottom left panel, but slightly differ, especially for PFS-values larger than 2, in the bottom middle and bottom right panels.

Results: Type I Error Rates

The Type I error rates (and power) of the PFSs did not depend on whether the MLE (truncated between -4.0 and 4.0), WLE, or MAP was used as the ability estimate in the computations in this paper. Therefore, only the results for the MLEs will be reported from the simulation study. Figure 2 shows the Type I error rates of the five PFSs for all test lengths for significance levels of 1% and 5%. The title of each panel denotes the test length and the level of significance. In each panel, the true examinee ability is shown along the X-axis, and the Type I error rate is shown along the Y-axis. Note that the range of the Y-axis is the same in the three panels on the left (all corresponding to 1% significance level) and the same in the three panels on the right (all corresponding to 5% significance level). For each PFS, the 9 values of the Type I error rate (at true ability of -2.0, -1.5, -1.0,...,1.5, and 2.0), shown using a solid circle, solid square, hollow circle, hollow square, or

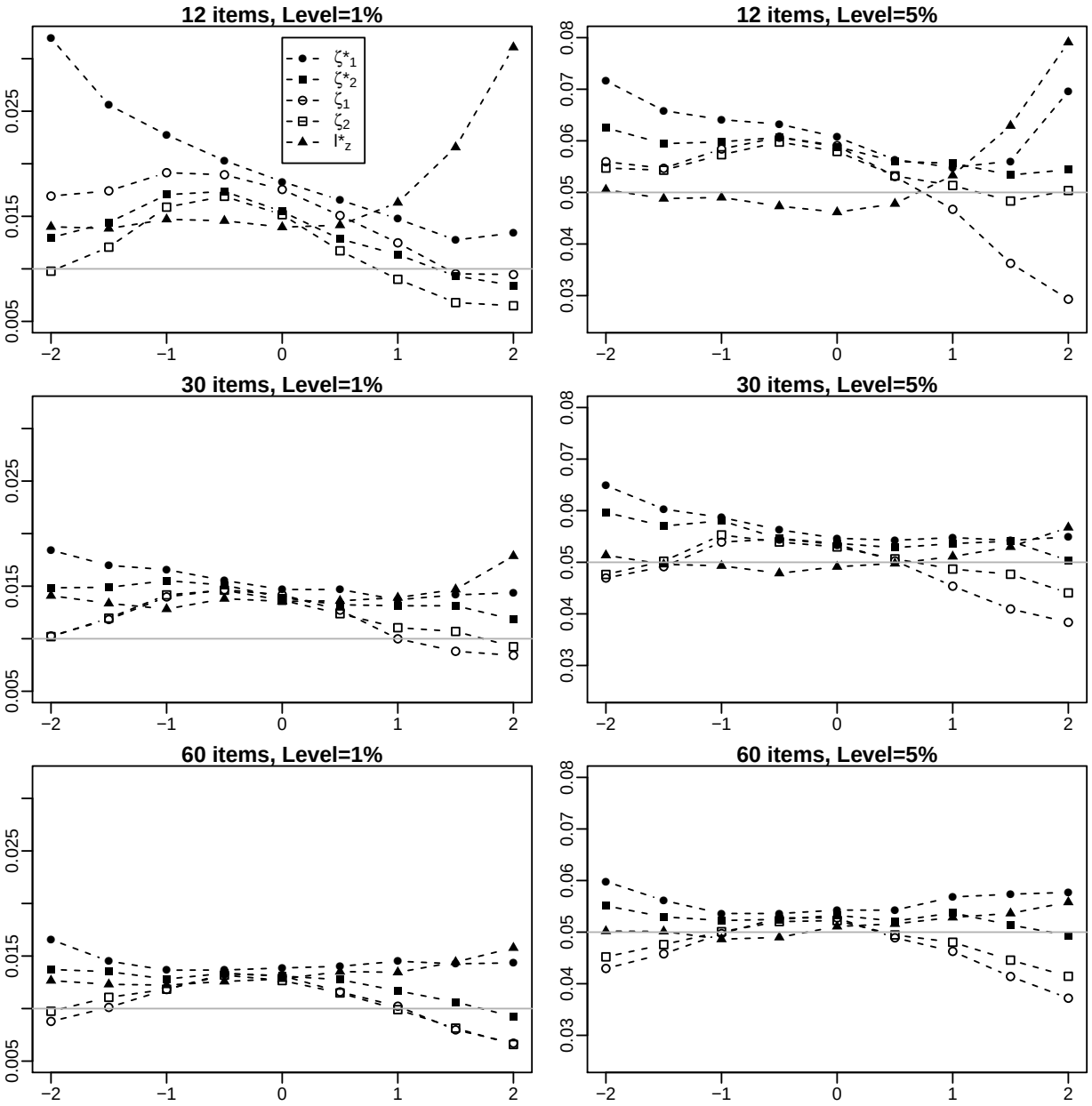

Figure 2. Type I Error Rates for all test lengths at 1% and 5% levels.

solid triangle (for ζ_1^*, ζ_2^*, ζ_1, ζ_2, and l_z^*, respectively), are joined using a dashed line in each panel. A horizontal solid line denotes the significance level in each panel.

Figure 2 shows that the Type I error rates of the PFSs are close to those of each other around true ability of 0. However, the Type I error rates of the PFSs become divergent as the true ability becomes extreme. Figure 2 further shows that:

- The relative performance of the PFSs at 1% level is very similar to that at 5% level
- The Type I error rates of ζ_1 are smaller than those of ζ_1^* and those of ζ_2 are smaller than those of ζ_2^*.
- The Type I error rates of ζ_2^* are smaller than those of ζ_1^*.
- For negative true abilities, the Type I error rates of l_z^* are closest to the nominal level among the three asymptotically corrected PFSs.
- For positive true abilities, the Type I error rates of ζ_2^* are closest to the nominal level among the three asymptotically corrected PFSs.
- The Type I error rates of l_z^* often exceed the significance level substantially, especially at 1% level and for large positive abilities. The inflation of the Type I error rates of l_z^* at small significance levels has been observed by several researchers such as Snijders (2001), van Krimpen-Stoop and Meijer (1999), and Sinharay (2016c).
- The Type I error rates of ζ_1^* and ζ_2^* are also inflated. Compared to l_z^*, the extent of Type I error inflation is slightly larger for ζ_1^* (thus, ζ_1^* has the largest Type I error rates on average among the PFSs considered here) and slightly smaller for ζ_2^*. The greatest extent of inflation is observed at extreme true abilities.
- The extent of inflation of Type I error rates is smaller at 5% level compared to 1% level.
- The Type I error rates of ζ_1 and ζ_2 are slightly larger than the nominal level around true ability of 0, but become smaller as true ability becomes more extreme, and are sometimes smaller than the nominal level for extreme true ability.

- For 60-item tests and 5% level, the Type I error rates of all the PFSs are smaller than 0.06 for all true abilities.

Results: Distribution for Short Tests

The Type I error rates of ζ_1^* and l_z^*, respectively, were found to rise sharply, especially at 1% level, for 12-item tests in Figure 2 for true ability of -2 and 2, respectively. Figure 3 shows the distributions of the observed values of ζ_1^*, ζ_2^*, and $-l_z^*$ for all the response patterns that fit the IRT model and were simulated using true abilities of -2, -1, 0, or 2 (shown in the title of each panel) for 12-item tests. The density of the standard normal distribution is also shown using a solid line. The figure shows that for 12-item tests, the null distributions of ζ_1^*, ζ_2^*, and l_z^* deviate from the standard normal distribution for all the true abilities and the deviation becomes larger as the true ability becomes more extreme. It seems that overall, the distribution of ζ_2^* is closer to the standard normal distribution compared to ζ_1^* and l_z^*—this is most clear in the bottom right panel where the distributions of ζ_1^* and l_z^* are much taller than that of the standard normal distribution. The small peaks for ability larger than 2 for ζ_1^* in the top left panel and $-l_z^*$ in the bottom right panel show that there are considerably more number of values of ζ_1^* and $-l_z^*$ larger than 2.0 in these cases than can be expected from a standard normal variable. This phenomenon results in their large Type I error rates in those cases (for true ability of -2 for ζ_1^* and 2 for l_z^*, both for 12-item tests). Also note that the empirical distribution of l_z^* for a test with six dichotomous items, shown in Figure 2B of Meijer and Tendeiro (2012), was considerably different from a standard normal distribution and looks very much like the bottom right panel of Figure 3 of this paper. However, Figure 2B of Meijer and Tendeiro (2012) or Figure 3 of this paper do not provide any evidence against the asymptotic normality of l_z^*, ζ_1^*, or ζ_2^*; the asymptotic normality of these PFSs holds for long tests, as noted by Meijer and Tendeiro (2012)—so it is natural that the normality did not hold for a 6-item test in Meijer and Tendeiro (2012) and for the 12-item test in this paper.

Even though the distribution of each of ζ_1^*, ζ_2^*, and l_z^* depart from a standard normal distribution for short tests, Figure 3 (and Figure 2 to a certain extent) shows that among

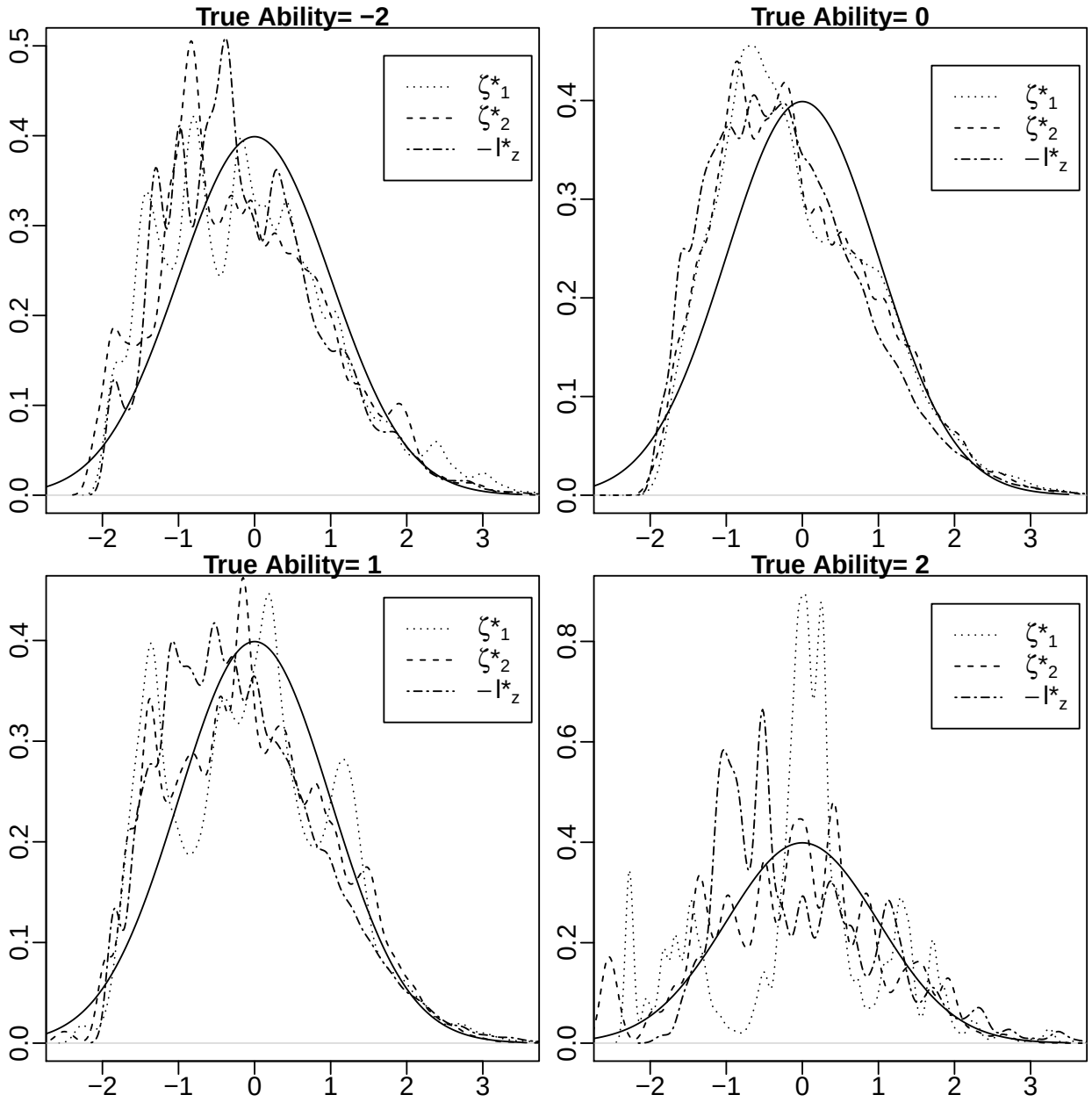

Figure 3. Null Distributions of ζ_1^*, ζ_2^*, and $-l_z^*$ for 12-item Tests.

these three PFS, the null distribution of ζ_2^* is closest to a standard normal distribution for short tests. Therefore, these figures show some evidence that ζ_2^* may be preferred as a PFS over l_z^* and ζ_1^* for short tests.

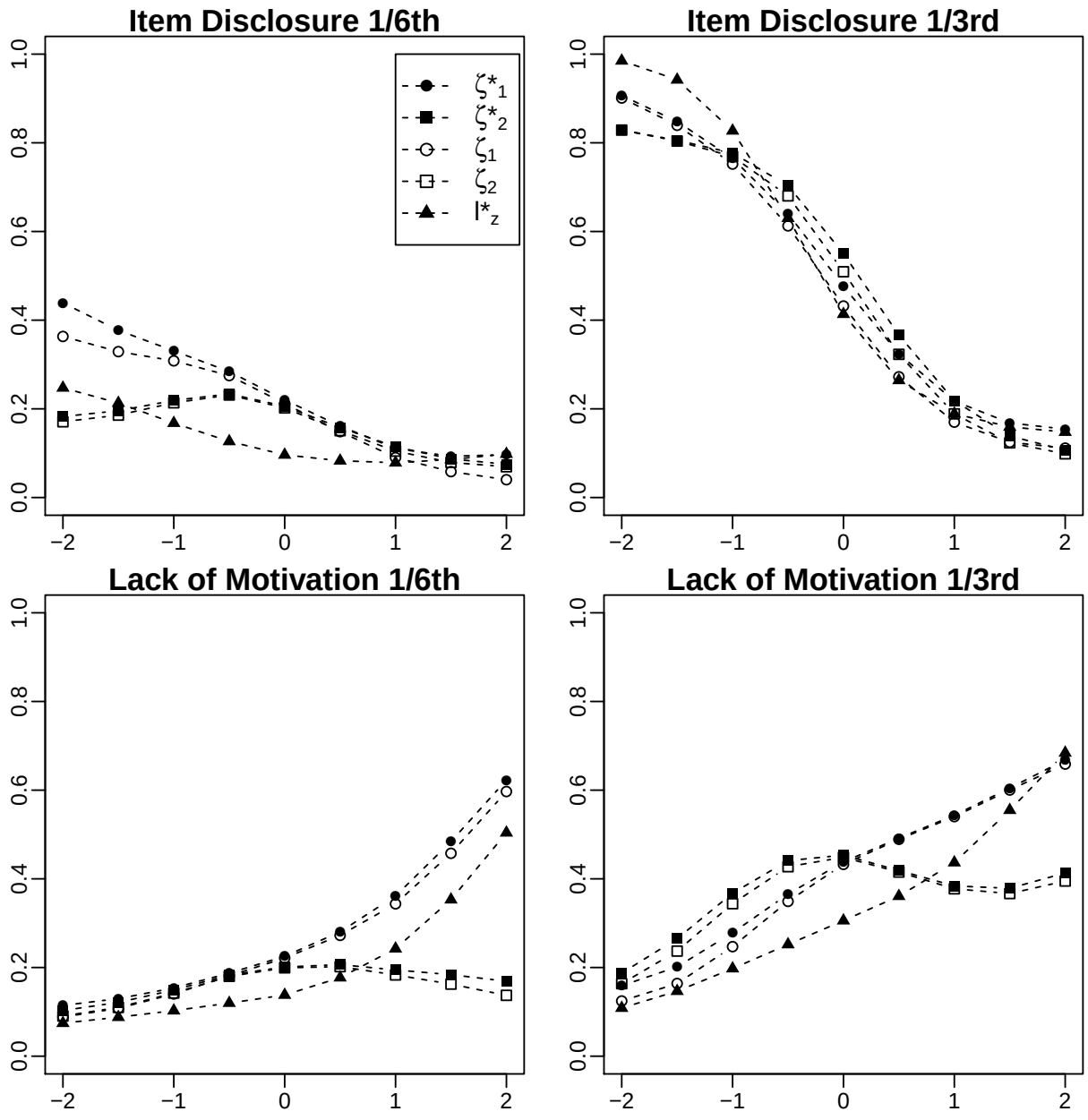

Figure 4. Power at 5% level for 12-item Tests.

Results: Power

Figures 4 to 6 show the values of power at 5% level for the different test lengths and different types of misfit. Each figure (representing a test length) shows the power under four types of misfit. The vertical axis in each panel of these figures ranges from 0 to 1. The figures show that ζ_1^* is always more powerful than ζ_1 and ζ_2^* is always more powerful than

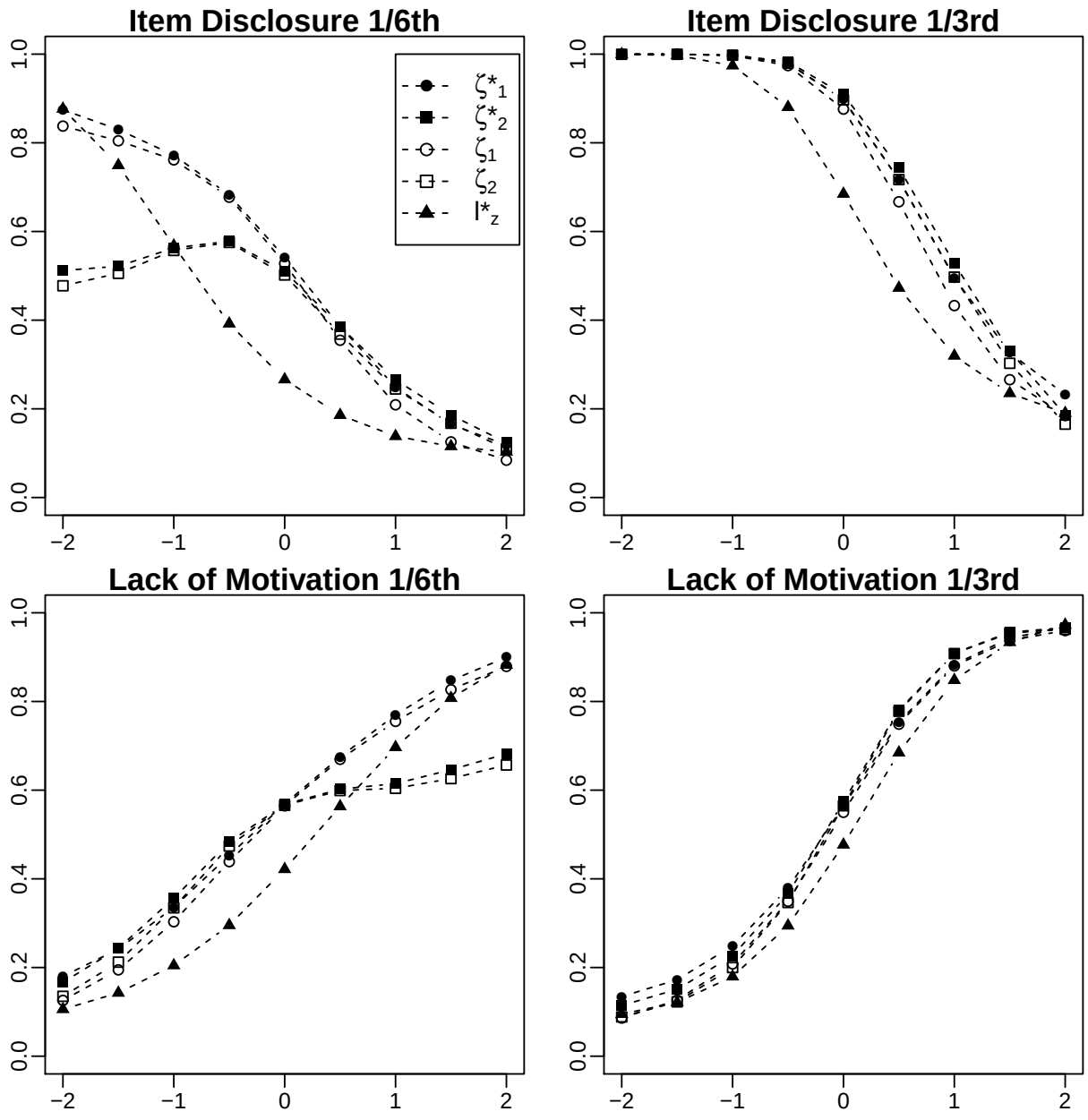

Figure 5. Power at 5% level for 30-item Tests.

ζ_2. In addition, these figures show that

- Either one among ζ_1^* and ζ_1 is more powerful than both of ζ_2^* and ζ_2 on some occasions, such as for $\theta < 0$ in the top left panel of Figure 4 and for $\theta > 0$ in the bottom left panel of Figure 4

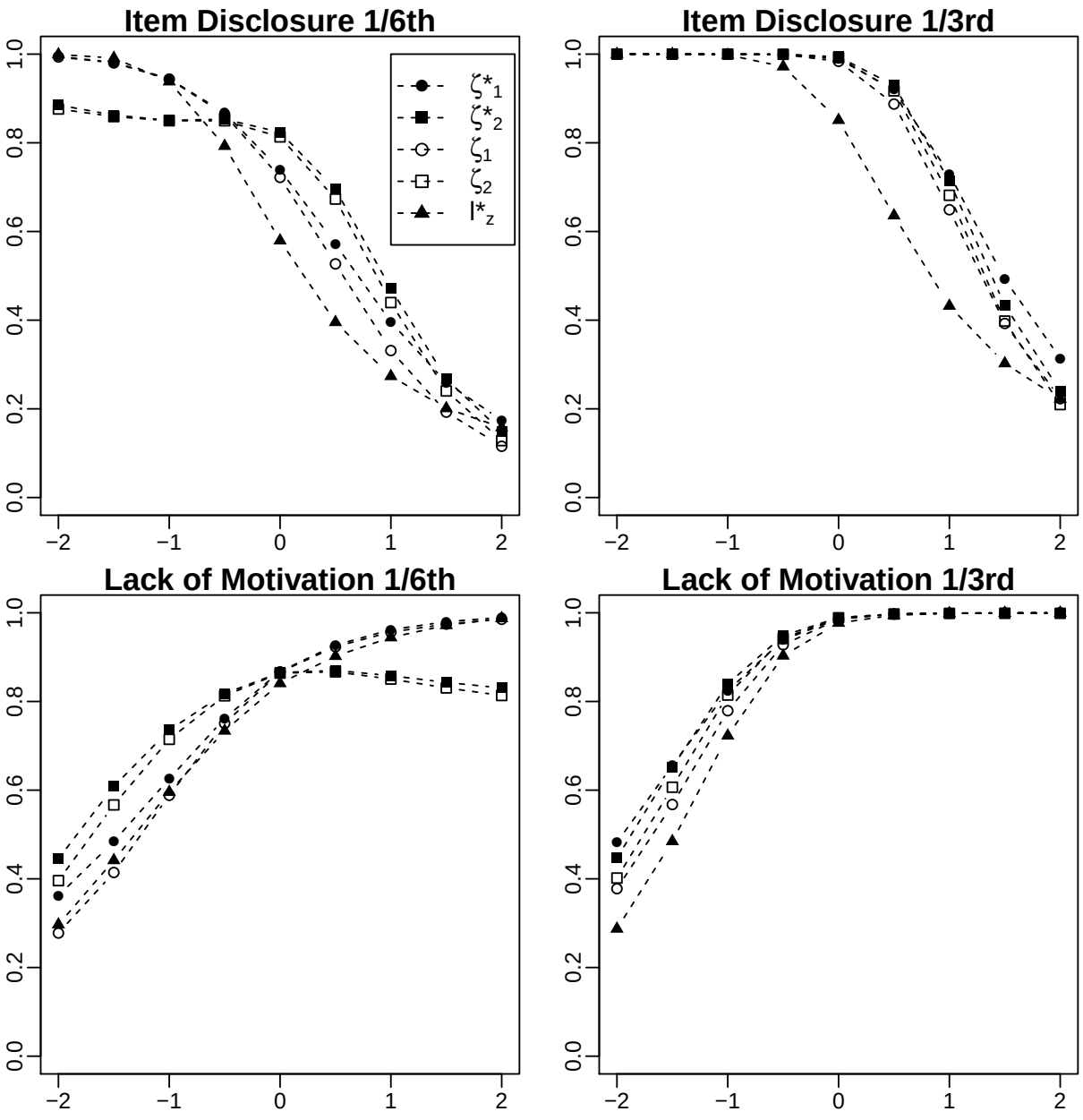

Figure 6. Power at 5% level for 60-item Tests.

- Either one among ζ_2^* and ζ_2 is more powerful than both of ζ_1^* and ζ_1 on some other occasions, such as for $-1 < \theta < 1$ in the top right panels of Figures 4 and 5

- The power of all of ζ_1^*, ζ_2^*, and ζ_2 are very close in the two rightmost panels of Figures 5 and 6, that is, when the test is at least moderately long and when the percentage of

aberrant responses is high

- The power of l_z^* is larger than all of ζ_1^*, ζ_1, ζ_2^*, and ζ_2 on a few occasions (e.g., for ability smaller than -1 in the top right panel of Figure 4), but is smaller than that of one of the others on most other occasions and is the smallest in several cases (such as around $\theta = 0$ in each panel in Figure 5).

Discussion of the Comparative Performance of the PFSs

The above results show that there is no single PFS that outperforms the others for all simulation conditions. For example, the Type I error rates of ζ_1^* are often inflated, but the power of ζ_1^* is often the largest; and the Type I error rates of l_z^* are often very close to the nominal level, but the power of l_z^* is often smaller than that of ζ_1^* and ζ_2^* on most occasions. This finding supports the statement of Tendeiro and Meijer (2014, p. 257) that different PFSs may have different sensibility to detect aberrant behavior under various testing conditions and probably imply that one should use all of ζ_1^*, ζ_2^*, and l_z^* to assess person fit for any given MFT and combine information from them (and potentially from other information sources) to make an overall decision on person fit for any given examinee.

However, the performance of ζ_2^* seems to be the best overall by a small margin. The Type I error rates of ζ_2^*, although slightly inflated occasionally (especially at 1% level), are the least inflated among those of ζ_1^*, ζ_2^*, and l_z^*; also, the Type I error rates of ζ_2^* at both 1% and 5% levels, though inflated, are almost always satisfactory according to Cochran's criterion for robustness (Cochran, 1952) that specifies that Type I error rates below 0.06 and 0.015 are satisfactory at 5% and 1% levels, respectively. Further, Figure 3 shows that for short tests, the distribution of ζ_2^* is closest to the theorized standard normal distribution among the three asymptotically standardized PFSs considered in this paper. Further, the power of ζ_2^* is often the largest, especially for the long tests, among all the PFSs considered in this paper.

Additional Simulations Where Item Parameters are Estimated

The above simulations were performed under the assumption that the true item parameters are known. This assumption is reasonable in several cases such as those with large samples for which accurate and precise estimates of item parameters are available and is common in existing simulation studies involving PFSs (e.g., Snijders, 2001; van Krimpen-Stoop & Meijer, 1999). However, in practice, true item parameters are never known and item parameters have to be estimated from the data. Therefore, an additional simulation was performed where item parameters were estimated. The simulation involved 1,000 repetitions of the following steps:

1. Simulate a set of true item parameters.

2. Simulate 1,000 true ability values (representing 1,000 examinees) for a data set uniformly from the nine above-mentioned values.

3. Use the above true item and ability parameters and the IRT model (3PLM+GPCM) to simulate a data set.

4. Estimate the item parameters using the marginal maximum likelihood algorithm from the data set.

5. Compute $\hat{\theta}_i$ of all the examinees from the data set using the estimated item parameters.

6. Compute ζ_1, ζ_2, ζ_1^*, ζ_2^*, and l_z^* for each examinee in the data set using $\hat{\theta}_i$ and the estimated item parameters. Compute the p-values corresponding to these PFSs under a standard normal null distribution assumption.

As in, for example, Glas and Meijer (2003), power was computed from data sets that included 90% examinees whose item-scores were simulated from the IRT model (3PLM+GPCM) and 10% examinees whose item-scores involved one type of misfit (lack of motivation or item disclosure).

The Type I error rates of the PFSs from this set of simulations were very similar to those in Figure 2 and are not reported here. The values of power of the PFSs from this set

of simulations are slightly smaller in general than those shown in Figures 4 to 6. Figure 7 shows the power at 5% level of the PFSs for 60-item tests when item parameters were estimated. The values of power in Figure 7 are mostly smaller than those in Figure 6,

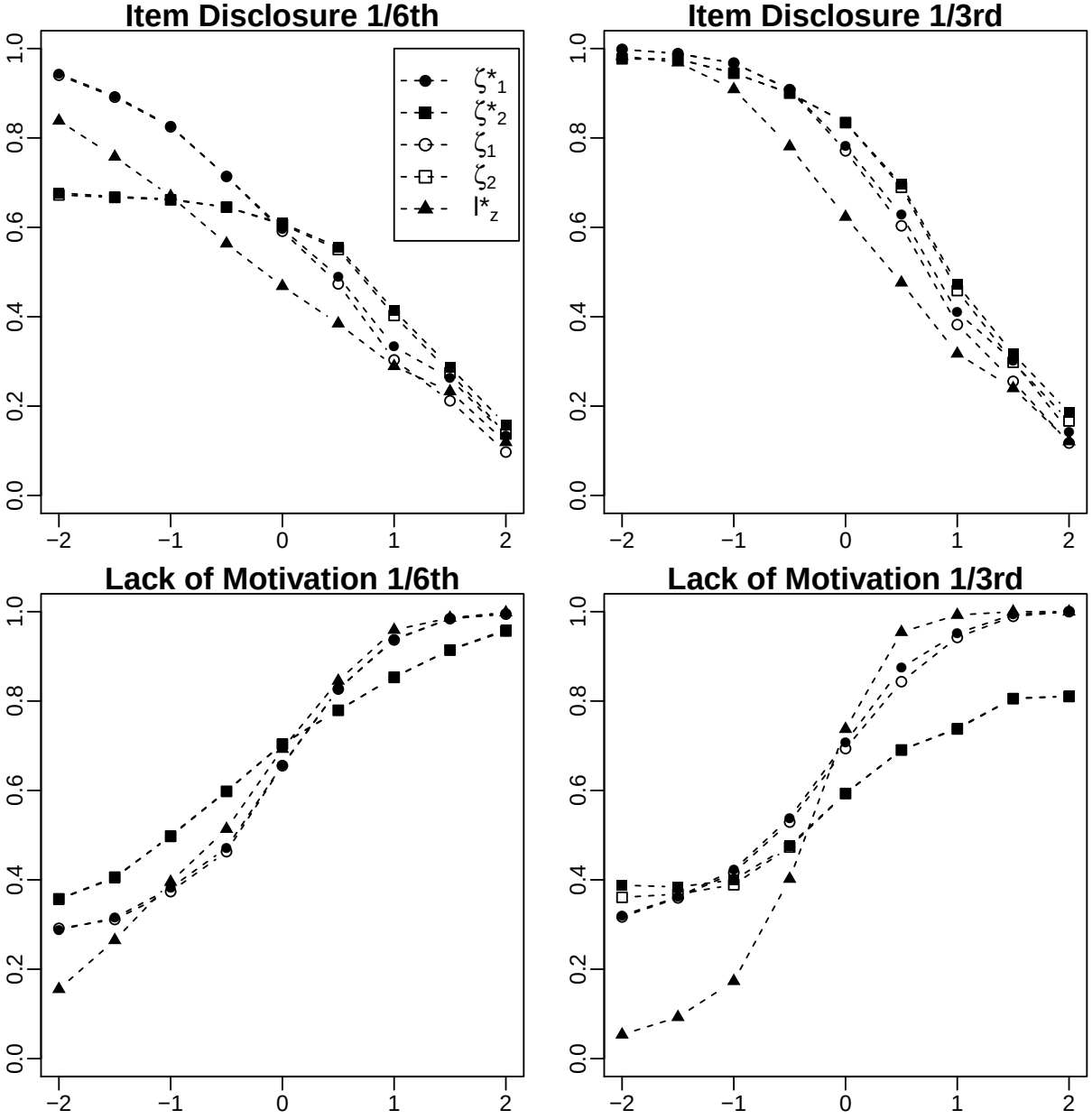

Figure 7. Power at 5% level for 60-item tests when item parameters were estimated.

especially in the top left and bottom right panels. For example, while the power of l_z^* for true ability of 0 is about 0.98 in the bottom right panel of Figure 6, the corresponding value

of power is about 0.74 in the bottom right panel of Figure 7. However, the comparative performance of the PFSs is similar in Figure 6 and Figure 7 in that ζ_1 and ζ_2 are less powerful than ζ_1^* and ζ_2^*, respectively, and it is difficult to pick an overall winner among the PFSs—no PFS has the largest power in all cases and each of the asymptotically standardized PFS has the largest power in some cases. The difference in power between ζ_1 and ζ_1^* (and also between ζ_2 and ζ_2^*) is smaller in Figure 7 than in Figure 6. Overall, the estimation of item parameters does not affect the comparative performance of the PFSs, especially of l_z^*, ζ_1^* and ζ_2^*.

Application to Real Data

Application 1: A Test with Only Polytomous Items

Let us consider the data set from NEO Personality Inventory that was analyzed by Glas and Dagohoy (2007) and Sinharay (2016c). [3] The NEO Personality Inventory is a personality test designed to provide a general description of normal personality that is relevant to clinical, counseling, and educational situations. The inventory consists of five broad domains. Each domain is measured by 48 items each of which is rated on a five-point scale. Data from 1,168 individuals on the neuroticism domain was analyzed in Glas and Dagohoy (2007) who split the 48 items on the domain into three sub-tests. The GPCM was fitted to each subtest separately and ζ_1, ζ_1^*, ζ_2, ζ_2^*, and l_z^* were computed for the examinees. The percent of individuals for which the PFSs were significant (under the assumption of a standard normal null distribution) at 5% significance level for the three subtests are provided in Table 2.

Table 2. The proportion of PFSs significant at 5% level for the first application.

Subtest	ζ_1	ζ_1^*	ζ_2	ζ_2^*	l_z^*
1	7.6	8.2	7.4	7.5	11.4
2	7.5	7.6	7.6	7.7	11.6
3	6.6	7.2	7.8	8.0	11.5

[3]The author is grateful to Cees A. W. Glas for generously sharing the data set.

The percent is largest for l_z^* for each subtest. However, the true person-misfit status is unknown for the data set—so it is impossible to state if the larger percent for l_z^* indicates its larger power or larger Type I error rate. The percent significant for each of ζ_1 and ζ_2 is slightly smaller than that of the corresponding asymptotically standardized PFS for each subtest.

Application 2: A Mixed-format Test

The PFSs were computed using data from one form of a licensure examination. The data set was analyzed in several chapters of Cizek and Wollack (2017). The test form includes 170 operational items that are dichotomously scored. Item scores on the form were available for 1,644 examinees. The licensure organization (that administers the examination) flagged, using a variety of statistical analysis and an investigative process that brought in other information, 48 individuals on the form as possible cheaters. The examinees flagged by the licensure organization can be considered truly aberrant for all practical purposes because of the rigorous nature of the flagging process of the organization.

Because the test included only dichotomous items, an (artificial) MFT was created by pooling 20 pairs of randomly chosen items into 20 3-category polytomous items and combining them with the 130 remaining items in the data set. Let's consider the polytomous item arising from a given item pair; an examinee is assigned a score of 0 if he/she answered both items in the pair incorrectly, is assigned a score of 1 if he/she answered exactly one item in the pair correctly, and is assigned a score of 2 if he/she answered both items in the pair correctly. The Rasch model is operationally used in the assessment; the model fitted to the resulting data set (with 130 dichotomous items and 20 three-category polytomous items) was a combination of the Rasch model for the dichotomous items and the partial credit model (Masters, 1982) for the polytomous items. The difficulty-parameters were estimated from the resulting data set and were used in the PFA. The MLE (truncated between -4.0 to 4.0) was used as the estimate of the examinee ability.

The proportions of examinees for which ζ_1, ζ_1^*, ζ_2, ζ_2^*, and l_z^* were significant at 5% significance level are provided in Table 3. The first row of numbers provides the proportions

Table 3. The proportion of PFSs significant at 5% level for the second application.

Examinees	ζ_1	ζ_1^*	ζ_2	ζ_2^*	l_z^*
All	0.08	0.09	0.08	0.08	0.11
Flagged	0.15	0.17	0.15	0.17	0.21

among all examinees. The second row of numbers provides the proportions among the 48 examinees who were flagged by the licensure organization; thus, for example, the proportion of 0.17 for ζ_1^* in the second row indicates that among the 48 examinees flagged by the licensure organization, ζ_1^* was significant for nine examinees (note that $8/48 \approx 0.17$).

Table 3 shows that the results for ζ_1, ζ_2, ζ_1^*, and ζ_2^* are very similar although the proportion significant among the flagged examinees for each of ζ_1, and ζ_2 is slightly smaller than of equal to that of the corresponding asymptotically standardized PFS. Table 3 also shows that the proportion significant for each PFS is about twice among the examinees flagged by the licensure organization compared to among all examinees—this result provides some evidence that the PFSs provide useful information.

Conclusions

This paper suggested four new PFSs—the extensions to MFTs of ζ_1 and ζ_2 (Tatsuoka, 1984) and their asymptotically corrected versions. The PFSs also apply to a test including only polytomous items. The asymptotically corrected versions were theoretically proved to follow the standard normal distribution asymptotically under no person misfit. In a simulation study, the Type I error rates of the corrected versions were found close to the nominal level at 5% level of significance, but slightly inflated at 1% level of significance. The corrected versions have slightly larger power than the corresponding un-corrected versions. In a real-data example where some score patterns were known to be aberrant, the PFSs provided useful information by flagging aberrant examinees at a larger rate than others. All these properties are in consonance with those for the existing asymptotically corrected PFSs (Magis, Beland, & Raiche, 2014; Snijders, 2001; Sinharay, 2016c, 2016b). Given the increasing importance of polytomous items (e.g., Darling-Hammond & Adamson,

2010) and PFA, the suggested PFSs promise to help practitioners.

The suggested PFSs are appropriate as PFSs when an investigator wants to test against an unspecified general alternative and may not be the most appropriate PFS for all applications of PFSs. For example, if one needs a PFS for a computerized-adaptive test, a PFS such as that of van Krimpen-Stoop and Meijer (2002) will be the most appropriate. If the anticipated model violation is more specific, a PFS such as the Lagrangian-multiplier statistic (Glas & Dagohoy, 2007) may be more powerful.

The suggested PFSs were used for PFA in this paper. However, they can be used for other purposes as well. For example, Tatsuoka (1984) used ζ_1 and ζ_2 for dichotomous items for diagnosing student misconceptions and all the suggested PFSs can be used for those purposes as well; ζ_1^* and ζ_2^* for MFTs, which were proved to follow an asymptotic standard normal null distribution, would help detect student misconceptions accurately.

There are several limitations of this paper and, consequently, several additional topics that can be investigated further. First, the Type I error rates of ζ_1^* and ζ_2^* were found to be larger than the nominal level for significance levels around 1%. Improvement on this Type I error inflation is a possible future research topic. It is possible to perform Monte Carlo simulations, as was performed by Sinharay (2016a) and van Krimpen-Stoop and Meijer (1999) for dichotomous items, to obtain an empirical null distribution of the PFSs. Second, Figure 2 shows that for short tests, the Type I error rates of ζ_1^* rapidly increases as the true ability becomes smaller than -1.5 and the Type I error rates of l_z^* rapidly increases as the true ability becomes larger than 1.5. It would be helpful to be able to explain this phenomenon. Third, it would be interesting to explore ability estimates other than the MLE, WLE, and MAP that are considered in this paper; robust ability estimates would be prime candidates because they would be less influenced by unusual responses, which may lead to larger power of the resulting PFSs. Sinharay (2016d) used robust ability estimates with asymptotically corrected PFSs for dichotomous items. Fourth, it may be of interest to examine the Type I error rate and power of the PFSs for more simulated and real data sets. For example, factors such as spread of item difficulties and average item discrimination often influence the power of PFSs in simulations (e.g., van Krimpen-Stoop

& Meijer, 1999); however, these factors were not manipulated in the simulations here and can be manipulated in future research. Finally, further research such as Conijn, Emons, and Sijtsma (2014), Meijer, Niessen, and Tendeiro (2015), and Meijer and Tendeiro (2014) should explore how PFA can be used in practice, for example, in high-stakes educational tests.

References

Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*, 318–338. (doi=10.1111/j.1745-3984.2010.00116.x)

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Washington, DC: Routledge.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315–345. (doi=10.1214/aoms/1177729380)

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

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning* (Tech. Rep.). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59–79. (doi=10.1177/014662168701100105)

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*, 67–86. (doi=10.1111/j.2044-8317.1985.tb00817.x)

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247. (doi=10.1177/0146621607302479)

Finkelman, M., & Kim, W. (2007, April). *Using person fit in a body of work standard setting.* Paper presented at the Annual meeting of the American Education Research Association, Chicago, IL.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180. (doi=10.1007/s11336-003-1081-5)

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233. (doi=10.1177/0146621603027003003)

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298. (doi=10.1207/s15324818ame1604_2)

Li, M. F., & Olenik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*, 215-231. (doi=10.1177/01466216970213002)

Lissitz, R. W., Hou, X., & Slater, S. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*, 1–50.

Magis, D., Beland, S., & Raiche, G. (2014). Snijders's correction of infit and outfit indexes with estimated ability level: An analysis with the Rasch model. *Journal of Applied Measurement*, *15*, 82–93.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. (doi=10.1007/bf02296272)

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2015). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52–62. (doi=10.1177/1073191115577800)

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. (doi=10.1177/01466210122031957)

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the l_z and l_z^* person-fit statistics and

problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*, 758–766. (doi=10.3102/1076998612466144)

Meijer, R. R., & Tendeiro, J. N. (2014). *The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us* (LSAC Research Report Series). Newtown, PA: Law School Admission Council.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75–106. (doi=10.1007/bf02294745)

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176. (doi=10.1002/j.2333-8504.1992.tb01436.x)

Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test securities irregularities.* Washington DC: Council of Chief State School Officers.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. (doi=10.1007/bf03372160)

Sijtsma, K. (1998). Methodological review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*, 3–32. (doi=10.1177/01466216980221001)

Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, *40*, 343–365. (doi=10.3102/1076998615589128)

Sinharay, S. (2016a). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement*, *53*, 63–85. (doi=10.1111/jedm.12101)

Sinharay, S. (2016b). Asymptotic Corrections of standardized extended caution indices. *Applied Psychological Measurement*, *40*, 418–433. (doi=10.1177/0146621616649963)

Sinharay, S. (2016c). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*, 992–1013. (doi=10.1007/s11336-015-9465-x)

Sinharay, S. (2016d). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of Mathematical and Statistical Psychology*, *69*, 175–193. (doi=10.1111/bmsp.12067)

Snijders, T. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342. (doi=10.1007/bf02294437)

Tao, J., Shi, N., & Chang, H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics*, *37*, 298-315. (doi=10.3102/1076998610393969)

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95-110. (doi=10.1007/bf02294208)

Tendeiro, J. N. (2017). The $l^*_{z(p)}$ person-fit statistic in an unfolding model context. *Applied Psychological Measurement*, *41*, 44–59. (doi=10.1177/0146621616669336)

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*, 239–259. (doi=10.1111/jedm.12046)

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327–345. (doi=10.1177/01466219922031446)

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*, 164–180. (doi=10.1177/01421602026002004)

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. (doi=10.1007/bf02294798)

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. (doi=10.1007/bf02294627)

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis [Computer Software].* Chicago, IL: Mesa Press.