PSYCHOMETRIKA—VOL. 89, NO. 2, 569–591
JUNE 2024
https://doi.org/10.1007/s11336-024-09960-x

EFFICIENT CORRECTIONS FOR STANDARDIZED PERSON-FIT STATISTICS

KYLIE GORNEY

MICHIGAN STATE UNIVERSITY

SANDIP SINHARAY AND CAROL ECKERLY

EDUCATIONAL TESTING SERVICE

Many popular person-fit statistics belong to the class of standardized person-fit statistics, T, and are assumed to have a standard normal null distribution. However, in practice, this assumption is incorrect since T is computed using (a) an estimated ability parameter and (b) a finite number of items. Snijders (Psychometrika 66(3):331–342, 2001) developed mean and variance corrections for T to account for the use of an estimated ability parameter. Bedrick (Psychometrika 62(2):191–199, 1997) and Molenaar and Hoijtink (Psychometrika 55(1):75–106, 1990) developed skewness corrections for T to account for the use of a finite number of items. In this paper, we combine these two lines of research and propose three new corrections for T that simultaneously account for the use of an estimated ability parameter and the use of a finite number of items. The new corrections are efficient in that they only require the analysis of the original data set and do not require the simulation or analysis of any additional data sets. We conducted a detailed simulation study and found that the new corrections are able to control the Type I error rate while also maintaining reasonable levels of power. A real data example is also included.

Key words: Person fit, item response theory, aberrant behavior.

Person-fit statistics are used to identify individuals who are displaying aberrant—or unusual—behavior. Many of the most popular person-fit statistics—including l_z (Drasgow et al. 1985), ζ_1 and ζ_2 (Tatsuoka 1984)—belong to the class of standardized person-fit statistics, T, and are assumed to have a standard normal null distribution. However, this distribution only holds when both of the following conditions are satisfied: (a) the true ability is known and is used to compute T and (b) an infinite number of items are available and are used to compute T. Numerous researchers have shown that when one or both of these conditions are not satisfied, the null distribution of T deviates from the standard normal distribution (e.g., Li & Olejnik 1997; Molenaar & Hoijtink 1990; Noonan et al. 1992; Reise 1995; Sinharay 2016b; Snijders 2001; van Krimpen-Stoop & Meijer 1999). Thus, in practical settings where both conditions are not satisfied (because the ability parameter is estimated and only a finite number of items are available), the assumption of a standard normal null distribution is incorrect and may lead to an inaccurate assessment of person fit. The person-fit assessment may be too liberal (resulting in an inflated Type I error rate), too conservative (resulting in an unnecessary sacrifice in power), or some combination of both.

Several corrections have been suggested to improve the accuracy of person-fit assessment when one or both of the above-mentioned conditions are not satisfied. Researchers such as de la Torre & Deng (2008), Glas & Meijer (2003), Sinharay (2016a), van Krimpen-Stoop & Meijer (1999) proposed resampling-based methods that simultaneously account for the use of an estimated ability parameter and the use of a finite number of items. However, resampling-based

Correspondence should be made to Kylie Gorney, Department of Counseling, Educational Psychology, and Special Education, Michigan State University, 460 Erickson Hall, 620 Farm Lane, East Lansing, MI 48824, USA. Email: kgorney@msu.edu

methods require the simulation and analysis of several large data sets and are therefore computationally intensive. More efficient methods have been proposed by Magis et al. (2014), Sinharay (2016b), Snijders (2001), who developed mean and variance corrections to account for the use of an estimated ability parameter, as well as Bedrick (1997) and Molenaar & Hoijtink (1990), who developed skewness corrections to account for the use of a finite number of items. These methods are efficient in that they only require the analysis of the original data set and do not require the simulation or analysis of any additional data sets. Notably, however, no efficient methods have been developed that simultaneously account for the use of an estimated ability parameter and the use of a finite number of items. The purpose of this paper is to fill this void in the literature.

In Sect. 1, we review the class of standardized person-fit statistics, T, as well as the existing corrections for T that account for either the use of an estimated ability parameter or the use of a finite number of items. In Sect. 2, we introduce three new corrections for T that simultaneously account for the use of an estimated ability parameter and the use of a finite number of items. All three corrections are computationally efficient. In Sect. 3, detailed simulations are conducted to (a) examine the null distributions and (b) compare the Type I error rates and power of the new and existing statistics. In Sect. 4, a real data example is provided. Finally, in Sect. 5, we conclude with a brief discussion and suggest directions for future research.

1. Background

Consider a test comprised of n items. Let X_i denote the score on item i, and let $p_i(\theta) = P(X_i = 1|\theta)$ denote the probability that item i is answered correctly given the ability parameter θ. For example, for the three-parameter logistic model (3PLM),

$$p_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{1}$$

where a_i, b_i, and c_i are the discrimination, difficulty, and pseudo-guessing parameters, respectively, of item i.

1.1. Standardized Person-Fit Statistics

Consider the class of standardized person-fit statistics that was introduced by Snijders (2001) and takes the form

$$T(\theta) = \frac{W(\theta)}{\sqrt{\mathrm{Var}(W(\theta))}}, \tag{2}$$

where

$$W(\theta) = \sum_{i=1}^{n}(X_i - p_i(\theta))w_i(\theta) \tag{3}$$

for some suitable weight function $w_i(\theta)$. For the standardized log-likelihood statistic l_z (Drasgow et al. 1985), the weight function is given by

$$w_i(\theta) = \log\frac{p_i(\theta)}{q_i(\theta)}, \tag{4}$$

where $q_i(\theta) = 1 - p_i(\theta)$. For the standardized extended caution indices ζ_1 and ζ_2 (Tatsuoka 1984), the weight functions are given by

$$w_i(\theta) = g - g_i \text{ and } w_i(\theta) = h(\theta) - p_i(\theta), \tag{5}$$

respectively, for

$$\begin{aligned}
g_i &= \frac{1}{N}\sum_{v=1}^{N} p_i(\theta_v),\\
h(\theta_v) &= \frac{1}{n}\sum_{i=1}^{n} p_i(\theta_v), \text{ and}\\
g &= \frac{1}{n}\sum_{i=1}^{n} g_i = \frac{1}{N\times n}\sum_{v=1}^{N}\sum_{i=1}^{n} p_i(\theta_v) = \frac{1}{N}\sum_{v=1}^{N} h(\theta_v),
\end{aligned}$$

where θ_v is the ability parameter of examinee v, and N is the total number of examinees.

Equation 3 implies that

$$\begin{aligned}
E(W(\theta)) &= \mu(\theta),\\
\text{Var}(W(\theta)) &= E\left[(W(\theta) - \mu(\theta))^2\right] = \sigma^2(\theta),\\
\text{Skew}(W(\theta)) &= E\left[\left(\frac{W(\theta)-\mu(\theta)}{\sigma(\theta)}\right)^3\right] = \gamma(\theta), \text{ and}\\
\text{Kurt}(W(\theta)) &= E\left[\left(\frac{W(\theta)-\mu(\theta)}{\sigma(\theta)}\right)^4\right] = \kappa(\theta),
\end{aligned}$$

where

$$\mu(\theta) = 0, \tag{6}$$

$$\sigma^2(\theta) = \sum_{i=1}^{n} p_i(\theta)q_i(\theta)w_i^2(\theta), \tag{7}$$

$$\gamma(\theta) = \frac{\sum_{i=1}^{n} p_i(\theta)q_i(\theta)(q_i(\theta) - p_i(\theta))w_i^3(\theta)}{\sigma^3(\theta)}, \text{ and} \tag{8}$$

$$\kappa(\theta) = \frac{\sum_{i=1}^{n} p_i(\theta)q_i(\theta)(1 - 3p_i(\theta)q_i(\theta))w_i^4(\theta)}{\sigma^4(\theta)}. \tag{9}$$

Therefore, the standardized person-fit statistic of Eq. 2 can be expressed as

$$T(\theta) = \frac{W(\theta) - \mu(\theta)}{\sigma(\theta)}. \tag{10}$$

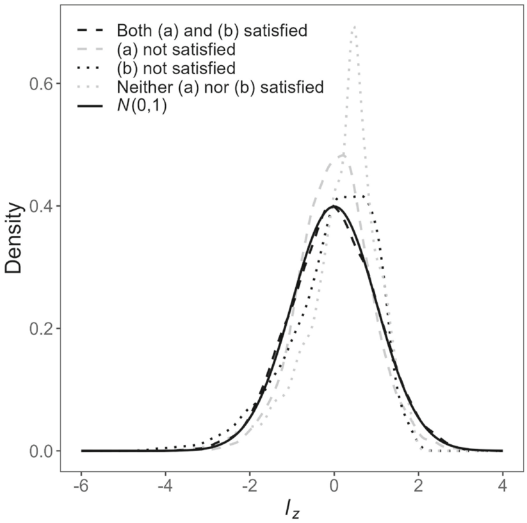

FIGURE 1.
The null distributions of l_z.

To classify a score pattern as aberrant, the significance probability p is computed as the probability that under the null distribution, the value of the test statistic T is equal to or exceeds the observed value t. That is,

$$p = \begin{cases} P(T \leq t) & \text{if extreme negative values of } t \text{ indicate misfit,} \\ P(T \geq t) & \text{if extreme positive values of } t \text{ indicate misfit.} \end{cases} \tag{11}$$

For the l_z statistic, extreme negative values indicate misfit. For the ζ_1 and ζ_2 statistics, extreme positive values indicate misfit.

Standardized person-fit statistics are assumed to have a standard normal null distribution. However, this distribution only holds when both of the following conditions are satisfied: (a) the true ability is known and is used to compute T and (b) an infinite number of items are available and are used to compute T. Figure 1 shows the null distribution of the l_z statistic when both conditions are (almost) satisfied—that is, when the true ability and 500 items are used to compute l_z. Observe that the null distribution (dashed black line) is very close to the theorized standard normal distribution (solid black line).

However, when condition (a) is not satisfied—that is, when the true ability is unknown and an ability estimate is used instead—the null distribution of T has a variance smaller than 1, as indicated by the dashed gray line in Fig. 1. Thus, the assumption of a standard normal null distribution (that has a variance of 1) leads to a conservative assessment of person fit. When condition (b) is not satisfied—that is, when a finite number of items are used to compute T—the null distribution of T is skewed, as indicated by the dotted black line in Fig. 1, which represents a 12-item test. The distribution is negatively skewed if extreme negative values of the statistic indicate misfit (e.g., l_z) or positively skewed if extreme positive values of the statistic indicate misfit (e.g., ζ_1, ζ_2). Thus, the assumption of a standard normal null distribution (that is not skewed) leads to a liberal assessment of person fit. When both (a) and (b) are not satisfied, the null distribution of T is skewed, has a variance smaller than 1, and has a mean that differs slightly from 0, as indicated by the dotted gray line in Fig. 1. Thus, the assumption of a standard normal null distribution leads to either a liberal or a conservative assessment of person fit, depending on the chosen significance level: the use of smaller significance levels leads to a more liberal assessment of person fit, while the use of larger significance levels leads to a more conservative assessment.

In an effort to obtain a more accurate assessment of person fit, several researchers have proposed corrections for T. Mean and variance corrections have been suggested to account for the use of an estimated ability parameter. Skewness corrections have been suggested to account for the use of a finite number of items. The following subsections contain reviews of each of these corrections.

1.2. Mean and Variance Corrections

When the true ability is unknown and an ability estimate is used instead, a naïve approximation of $T(\theta)$ can be obtained by inserting $\hat{\theta}$ into Eq. 10. That is,

$$T(\hat{\theta}) = \frac{W(\hat{\theta}) - \mu(\hat{\theta})}{\sigma(\hat{\theta})}. \tag{12}$$

However, Snijders (2001) proved that replacing θ with $\hat{\theta}$ has a non-negligible effect on the variance of W—and therefore, T—even when an infinite number of items are used. Thus, the assumption that $T(\hat{\theta})$ has a standard normal null distribution, even asymptotically, is incorrect. Snijders further showed that if $\hat{\theta}$ satisfies the condition

$$r_0(\hat{\theta}) + \sum_{i=1}^{n} (X_i - p_i(\hat{\theta}))r_i(\hat{\theta}) = 0 \tag{13}$$

for some functions $r_0(\hat{\theta})$ and $r_i(\hat{\theta})$, then the mean and variance of $W(\hat{\theta})$ can be approximated using

$$E(W(\hat{\theta})) \approx \tilde{\mu}(\hat{\theta}) \text{ and}$$
$$\text{Var}(W(\hat{\theta})) \approx \tilde{\sigma}^2(\hat{\theta}),$$

respectively, where

$$\tilde{\mu}(\hat{\theta}) = -c(\hat{\theta})r_0(\hat{\theta}) \tag{14}$$

and

$$\tilde{\sigma}^2(\hat{\theta}) = \sum_{i=1}^{n} p_i(\hat{\theta})q_i(\hat{\theta})\tilde{w}_i^2(\hat{\theta}) \tag{15}$$

for

$$c(\hat{\theta}) = \frac{\sum_{i=1}^{n} p_i'(\hat{\theta})w_i(\hat{\theta})}{\sum_{i=1}^{n} p_i'(\hat{\theta})r_i(\hat{\theta})}, \tag{16}$$

where $p_i'(\hat{\theta})$ is the first derivative of $p_i(\hat{\theta})$ with respect to $\hat{\theta}$, and the modified weight function is given by

$$\tilde{w}_i(\hat{\theta}) = w_i(\hat{\theta}) - c(\hat{\theta})r_i(\hat{\theta}). \tag{17}$$

Therefore, the asymptotically correct statistic $T^*(\hat{\theta})$ can be derived from $T(\hat{\theta})$ by adjusting both the mean and variance of the statistic. In other words,

$$T^*(\hat{\theta}) = \frac{W(\hat{\theta}) - \tilde{\mu}(\hat{\theta})}{\tilde{\sigma}(\hat{\theta})} \tag{18}$$

has an asymptotic standard normal null distribution.

The corrected statistic T^* can be computed using any ability estimate $\hat{\theta}$ that has functions $r_0(\hat{\theta})$ and $r_i(\hat{\theta})$ which satisfy Eq. 13. Magis et al. (2012) showed that for the weighted likelihood (WL) estimate (Warm 1989), maximum likelihood (ML) estimate, and maximum a posteriori (MAP) estimate, Eq. 13 is satisfied for

$$r_i(\hat{\theta}) = \frac{p_i'(\hat{\theta})}{p_i(\hat{\theta})q_i(\hat{\theta})}$$

and

$$r_0(\hat{\theta}) = \begin{cases} \frac{J(\hat{\theta})}{2I(\hat{\theta})} & \text{if } \hat{\theta} \text{ is the WL estimate,} \\ 0 & \text{if } \hat{\theta} \text{ is the ML estimate,} \\ \frac{d \log f(\hat{\theta})}{d\hat{\theta}} & \text{if } \hat{\theta} \text{ is the MAP estimate,} \end{cases}$$

where

$$J(\hat{\theta}) = \sum_{i=1}^{n} \frac{p_i'(\hat{\theta})p_i''(\hat{\theta})}{p_i(\hat{\theta})q_i(\hat{\theta})},$$

$$I(\hat{\theta}) = \sum_{i=1}^{n} \frac{[p_i'(\hat{\theta})]^2}{p_i(\hat{\theta})q_i(\hat{\theta})},$$

and $f(\cdot)$ is the prior distribution on θ.

1.3. Skewness Corrections

Researchers such as Molenaar & Hoijtink (1990), Noonan et al. (1992), and van Krimpen-Stoop & Meijer (1999) have shown that the null distribution of T becomes more skewed as test length decreases. Therefore, the standard normal distribution (that is not skewed) provides a poor approximation for tests with fewer items. To obtain a more accurate approximation of the null distribution of T, several methods have been developed that take this skewness into account. These methods use naïve approximations of the mean, variance, and skewness of $W(\hat{\theta})$ that are obtained by inserting $\hat{\theta}$ into Eqs. 6, 7, and 8, respectively. For example, the naïve approximation of skewness is given by

$$\gamma(\hat{\theta}) = \frac{\sum_{i=1}^{n} p_i(\hat{\theta})q_i(\hat{\theta})(q_i(\hat{\theta}) - p_i(\hat{\theta}))w_i^3(\hat{\theta})}{\sigma^3(\hat{\theta})}. \tag{19}$$

Molenaar & Hoijtink (1990) suggested two methods to approximate the null distribution of T. The first method is based on the Cornish–Fisher expansion. The skewness-corrected statistic is given by

$$T_{\mathrm{CF}}(\hat{\theta}) = T(\hat{\theta}) - \frac{\gamma(\hat{\theta})[(T(\hat{\theta}))^2 - 1]}{12}, \tag{20}$$

which is equivalent to approximating the significance probability of Eq. 11 as

$$p_{\mathrm{CF}} = \begin{cases} \Phi\left(T_{\mathrm{CF}}(\hat{\theta})\right) & \text{if extreme negative values of } t \text{ indicate misfit,} \\ 1 - \Phi\left(T_{\mathrm{CF}}(\hat{\theta})\right) & \text{if extreme positive values of } t \text{ indicate misfit,} \end{cases} \tag{21}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution.

The second method of Molenaar & Hoijtink (1990) employs a higher-order approximation of the significance probability that is based on a χ^2 distribution with $\nu(\hat{\theta}) = \frac{8}{\gamma^2(\hat{\theta})}$ degrees of freedom. Thus, the significance probability is approximated as

$$p_{\chi^2} = \begin{cases} P\left(\chi^2(\nu(\hat{\theta})) \geq \frac{|W(\hat{\theta}) - \mu(\hat{\theta}) - a(\hat{\theta})|}{b(\hat{\theta})}\right) & \text{if extreme negative values of } t \text{ indicate misfit,} \\ P\left(\chi^2(\nu(\hat{\theta})) \geq \frac{|W(\hat{\theta}) - \mu(\hat{\theta}) + a(\hat{\theta})|}{b(\hat{\theta})}\right) & \text{if extreme positive values of } t \text{ indicate misfit,} \end{cases} \tag{22}$$

where

$$a(\hat{\theta}) = b(\hat{\theta})\nu(\hat{\theta}) \text{ and}$$
$$b(\hat{\theta}) = \sqrt{\frac{\sigma^2(\hat{\theta})}{2\nu(\hat{\theta})}}.$$

A third method was suggested by Bedrick (1997), who used the Edgeworth expansion to approximate the significance probability as

$$p_{\mathrm{EW}} = \begin{cases} \Phi(T(\hat{\theta})) - \frac{\phi(T(\hat{\theta}))\gamma(\hat{\theta})[(T(\hat{\theta}))^2 - 1]}{6} & \text{if extreme negative values of } t \text{ indicate misfit,} \\ 1 - \left(\Phi(T(\hat{\theta})) - \frac{\phi(T(\hat{\theta}))\gamma(\hat{\theta})[(T(\hat{\theta}))^2 - 1]}{6}\right) & \text{if extreme positive values of } t \text{ indicate misfit,} \end{cases} \tag{23}$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal distribution. The Edgeworth expansion occasionally yields estimates of p that are smaller than 0 or larger than 1. In this paper, we replace such values with (traditional) estimates of p that are obtained by applying $T(\hat{\theta})$ under the assumption of a standard normal null distribution.

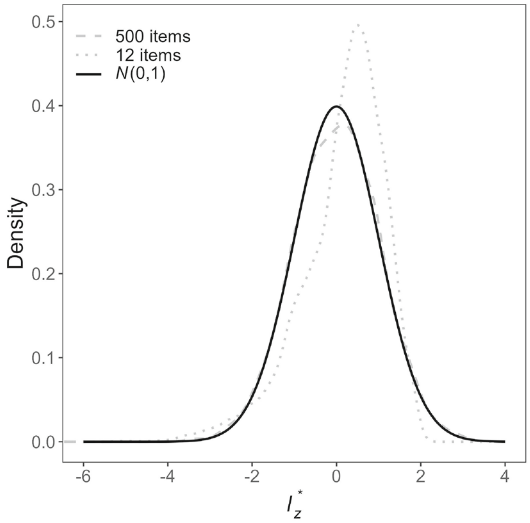

Figure 2.
The null distributions of l_z^*.

We note that it may be desirable to transform the significance probability approximations in Eqs. 22 and 23 to person-fit statistics that are approximately standard normally distributed. If extreme negative values of t indicate misfit, the significance probability approximations can be transformed using the inverse CDF of the standard normal distribution, $\Phi^{-1}(p)$. If extreme positive values of t indicate misfit, the significance probability approximations can be transformed using $\Phi^{-1}(1-p)$.

2. Method

In the previous section, we reviewed (a) the class of standardized person-fit statistics T, (b) mean and variance corrections for T that account for the use of an estimated ability parameter, and (c) skewness corrections for T that account for the use of a finite number of items. In this section, we apply mean, variance, and skewness corrections to simultaneously account for the use of an estimated ability parameter and the use of a finite number of items.

We start by considering the class of mean and variance-corrected statistics T^* that is given by Eq. 18. As was the case for the uncorrected statistic T, the corrected statistic T^* is assumed to have a standard normal null distribution. However, this distribution only holds asymptotically—that is, when an infinite number of items are available. Figure 2 shows the null distribution of the l_z^* statistic when 500 items are used to compute l_z^*. Observe that the null distribution (dashed gray line) is very close to the theorized standard normal distribution (solid black line). Yet, when a finite number of items are used to compute T^*, the null distribution of T^* is skewed, as indicated by the dotted gray line in Fig. 2. Thus, the assumption of a standard normal distribution (that is not skewed) leads to a liberal assessment of person fit, especially at smaller significance levels such as $\alpha = .01$ or $\alpha = .02$ (e.g., de la Torre & Deng 2008; Sinharay 2016b; Snijders 2001).

In an effort to obtain a more accurate assessment of person fit, we introduce three new methods for approximating the null distribution of T^* that take this skewness into account. These methods use asymptotic approximations of the mean, variance, and skewness of $W(\hat{\theta})$ that are given by

Eqs. 14, 15, and

$$\tilde{\gamma}(\hat{\theta}) = \frac{\sum_{i=1}^{n} p_i(\hat{\theta}) q_i(\hat{\theta})(q_i(\hat{\theta}) - p_i(\hat{\theta}))\tilde{w}_i^3(\hat{\theta})}{\tilde{\sigma}^3(\hat{\theta})}, \tag{24}$$

respectively.

The three new methods for approximating the null distribution of T^* are heuristics that parallel the methods in Eqs. 21, 22, and 23 for approximating the null distribution of T. R code to implement the new methods is included in "Appendix A".

The first of the three new methods is based on the Cornish–Fisher expansion. The skewness-corrected statistic is given by

$$T^*_{\mathrm{CF}}(\hat{\theta}) = T^*(\hat{\theta}) - \frac{\tilde{\gamma}(\hat{\theta})[(T^*(\hat{\theta}))^2 - 1]}{12}, \tag{25}$$

which is equivalent to approximating the significance probability of Eq. 11 as

$$p^*_{\mathrm{CF}} = \begin{cases} \Phi\left(T^*_{\mathrm{CF}}(\hat{\theta})\right) & \text{if extreme negative values of } t^* \text{ indicate misfit,} \\ 1 - \Phi\left(T^*_{\mathrm{CF}}(\hat{\theta})\right) & \text{if extreme positive values of } t^* \text{ indicate misfit.} \end{cases} \tag{26}$$

The second method employs a higher-order approximation of the significance probability that is based on a χ^2 distribution with $\tilde{\nu}(\hat{\theta}) = \frac{8}{\tilde{\gamma}^2(\hat{\theta})}$ degrees of freedom. Thus, the significance probability is approximated as

$$p^*_{\chi^2} = \begin{cases} P\left(\chi^2(\tilde{\nu}(\hat{\theta})) \geq \frac{|W(\hat{\theta}) - \tilde{\mu}(\hat{\theta}) - \tilde{a}(\hat{\theta})|}{\tilde{b}(\hat{\theta})}\right) & \text{if extreme negative values of } t^* \text{ indicate misfit,} \\ P\left(\chi^2(\tilde{\nu}(\hat{\theta})) \geq \frac{|W(\hat{\theta}) - \tilde{\mu}(\hat{\theta}) + \tilde{a}(\hat{\theta})|}{\tilde{b}(\hat{\theta})}\right) & \text{if extreme positive values of } t^* \text{ indicate misfit,} \end{cases} \tag{27}$$

where

$$\tilde{a}(\hat{\theta}) = \tilde{b}(\hat{\theta})\tilde{\nu}(\hat{\theta}) \text{ and}$$
$$\tilde{b}(\hat{\theta}) = \sqrt{\frac{\tilde{\sigma}^2(\hat{\theta})}{2\tilde{\nu}(\hat{\theta})}}.$$

The third method uses the Edgeworth expansion to approximate the significance probability as

$$p^*_{\mathrm{EW}} = \begin{cases} \Phi(T^*(\hat{\theta})) - \frac{\phi(T^*(\hat{\theta}))\tilde{\gamma}(\hat{\theta})[(T^*(\hat{\theta}))^2 - 1]}{6} & \text{if extreme negative values of } t^* \text{ indicate misfit,} \\ 1 - \left(\Phi(T^*(\hat{\theta})) - \frac{\phi(T^*(\hat{\theta}))\tilde{\gamma}(\hat{\theta})[(T^*(\hat{\theta}))^2 - 1]}{6}\right) & \text{if extreme positive values of } t^* \text{ indicate misfit.} \end{cases} \tag{28}$$

If the Edgeworth expansion yields estimates of p that are smaller than 0 or larger than 1, we replace such values with estimates of p that are obtained by applying $T^*(\hat{\theta})$ under the assumption of a standard normal null distribution.

The inverse CDF method can be used to transform the significance probability approximations in Eqs. 27 and 28 to person-fit statistics that are approximately standard normally distributed. If extreme negative values of t^* indicate misfit, the significance probability approximations can be transformed using $\Phi^{-1}(p^*)$. If extreme positive values of t^* indicate misfit, the significance probability approximations can be transformed using $\Phi^{-1}(1 - p^*)$.

3. Simulation Study

3.1. Design and Analysis

We conducted a detailed simulation study to (a) examine the null distributions and (b) compare the Type I error rates and power of the new and existing statistics. The simulations were designed to mimic realistic testing conditions—therefore, an estimated ability parameter and a finite number of items were used to compute the person-fit statistics.

Three test lengths (12, 36, 72) were studied to represent short, medium, and long tests. For each test length, 1 million (10,000 examinees × 100 replications) score patterns were simulated. For each replication, new sets of person and item parameters were generated. As in Glas & Meijer (2003), 90% of the examinees were non-aberrant—that is, they fit the model—and were used to study the null distributions and Type I error rates of the statistics. The remaining 10% of the examinees were divided equally into four groups of aberrant examinees and were used to study power. The four groups of aberrant examinees were characterized by the type of aberrant behavior (lack of motivation, item disclosure) and by the proportion of contaminated items $(\frac{1}{6}, \frac{1}{3})$.

Uncontaminated item scores were generated using the 3PLM. For each replication, the item parameters were sampled such that $a_i \sim Lognormal(0, 0.25^2)$, $b_i \sim \mathcal{N}(0, 1)$, and $c_i \sim \mathcal{U}(0.05, 0.30)$, as in Sinharay (2016b), and the person parameters were sampled such that $\theta_v \sim \mathcal{N}(0, 1)$. Contaminated item scores were generated after manipulating the item success probabilities. As in Glas & Meijer (2003), lack of motivation was simulated as random guessing on the easiest items, with a success probability equal to 0.2. Item disclosure was simulated as preknowledge of the most difficult items, with a success probability equal to 0.9. By simulating lack of motivation on the easiest items and item disclosure on the most difficult items, we studied conditions in which ability estimates would be severely impacted and are therefore important to detect.

After simulating the data, each of the score patterns was analyzed 72 times: once for each combination of two classes of person-fit statistics (T, T^*), three weight functions (l_z, ζ_1, ζ_2), four skewness corrections (none, Cornish–Fisher expansion, χ^2 approximation, Edgeworth expansion), and three ability estimates (WL, ML, MAP). In all conditions, the item parameters were treated as known. This assumption is common in person-fit research, as it prevents the null distribution of the statistics from being affected by any uncertainty in the item parameter estimates (e.g., Molenaar & Hoijtink 1990; Snijders 2001; van Krimpen-Stoop & Meijer 1999).

The ML and MAP estimates of ability were bounded between -4 and 4. The standard normal distribution was used as the prior distribution for the MAP estimates. The standardized extended caution indices were computed after reversing the sign of the weights given in Eq. 5; thus, extreme negative values of the statistics indicated misfit. This adjustment was made to facilitate comparisons against the standardized log-likelihood statistics, for which extreme negative values also indicate misfit.

3.2. Results

The choice of ability estimate was found not to affect the relative performance of the person-fit statistics. Therefore, we focus on the WL estimates of ability (since they were computed without any bounds or prior distributions) and include results for the other ability estimates in “Appendix B”.

3.2.1. The Null Distributions of the Person-Fit Statistics Figure 3 displays the first four moments of the null distributions of the person-fit statistics. Each row corresponds to a different moment (mean, variance, skewness, excess kurtosis), and each column corresponds to a different test length (12, 36, 72). Note that excess kurtosis is defined as the kurtosis minus 3. Horizontal dotted lines are used to indicate the values that are expected under the theoretical null distribution. It is desirable for the moments of the empirical null distributions to be as close to these values as possible.

Figure 3 reveals that the null distribution of T (i.e., the uncorrected statistic given by Eq. 12) is negatively skewed, has a variance smaller than 1, and has a mean that is slightly larger than 0. Similar results are shown in Table 1 of Li & Olejnik (1997), Table 3 of Reise (1995), and Table 1 of van Krimpen-Stoop & Meijer (1999). The skewness-corrected statistics T_{CF}, T_{χ^2}, and T_{EW} return values of skewness that are closer to 0, but do not offer much improvement in terms of the mean or variance. Conversely, the mean and variance-corrected statistic T^* (that is given by Eq. 18) has a variance that is closer to 1, but does not offer much improvement in terms of skewness. Interestingly, although l_z^* has a mean that is closer to 0, ζ_1^* and ζ_2^* have means that are farther from 0. Similar results are shown in Tables 1 and 2 of van Krimpen-Stoop & Meijer (1999) for l_z^*, and in Table 1 of Sinharay (2016b) for ζ_1^* and ζ_2^*.

The newly proposed statistics T_{CF}^*, $T_{\chi^2}^*$, and T_{EW}^* are the only statistics to incorporate mean, variance, and skewness corrections. Therefore, it is not surprising to see that these are the only statistics that improve both the skewness and the variance. Figure 3 also reveals that although these statistics have similar means and variances, $T_{\chi^2}^*$ is the most effective at reducing skewness, followed by T_{EW}^* and then T_{CF}^*. This finding parallels the results for the class of T statistics, as shown in Fig. 3 and in previous research (Bedrick 1997; Molenaar & Hoijtink 1990; Santos et al. 2020; von Davier & Molenaar 2003).

3.2.2. Type I Error Rates Figure 4 displays the Type I error rates of the person-fit statistics. Each row corresponds to a different significance level (.01, .02, .05, .10), and each column corresponds to a different test length (12, 36, 72). Horizontal dotted lines are used to indicate the significance levels. It is desirable for the Type I error rates to be at or below these lines.

Figure 4 reveals that the Type I error rates of T (i.e., the uncorrected statistic) vary depending on the weight function that is used. For l_z, the Type I error rate is consistently smaller than the nominal level. This result can largely be attributed to the reduced variance of the statistic (see Fig. 3). In contrast, for ζ_1 and ζ_2, the Type I error rates are slightly larger than the nominal level when α is small and the test is short, but are close to or smaller than the nominal level in all other instances.

The Type I error rates of the skewness-corrected statistics T_{CF}, T_{χ^2}, and T_{EW} are always smaller than the Type I error rates of T. While this result is desirable in instances where the Type I error rates of T are inflated (e.g., when ζ_1 and ζ_2 are used with a small α and a short test), it is undesirable in instances where T is already conservative, which seems to be the more common case. Therefore, T_{CF}, T_{χ^2}, and T_{EW} are of limited utility in the present context.

The Type I error rates of the mean and variance-corrected statistic T^* are always larger than the Type I error rates of T. When $\alpha = .10$, this result is useful, since the Type I error rates of T^* are closer to, but still do not exceed, the nominal level. In contrast, when $\alpha = .01$, .02, or .05, the Type I error rates of T^* often exceed the nominal level, which is undesirable in practice.

The Type I error rates of the newly proposed statistics T_{CF}^*, $T_{\chi^2}^*$, and T_{EW}^* are generally quite favorable and seem to overcome the limitations of the other corrected statistics. That is, the Type

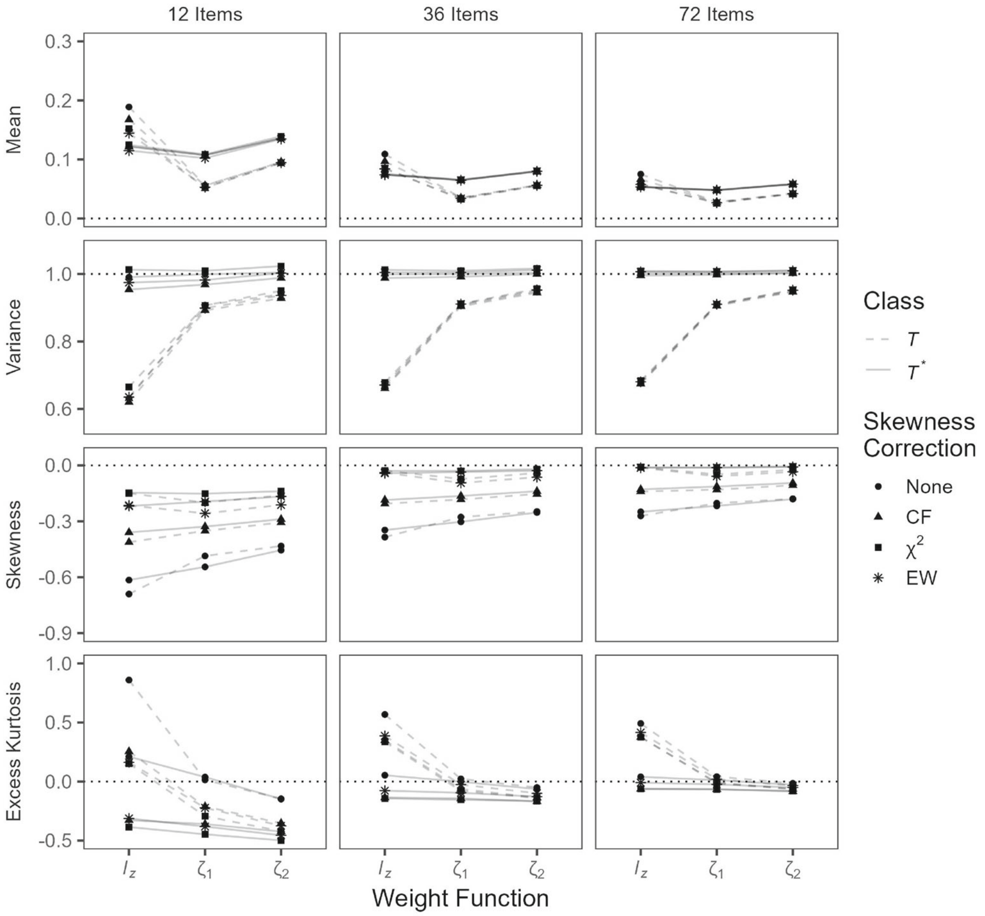

FIGURE 3.
Descriptive statistics of the null distributions of the person-fit statistics. *CF* Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

I error rates tend to be close to the nominal level (unlike T_{CF}, T_{χ^2}, and T_{EW}), while still not exceeding it (unlike T^*). Of the three new statistics, T^*_{CF} produces Type I error rates that are closest to the nominal level, followed by T^*_{EW} when $\alpha = .01$, or $T^*_{\chi^2}$ when $\alpha = .02, .05$, or $.10$. These conclusions are the same regardless of test length or the choice of weight function.

Figure 5 displays the Type I error rates by quintile of the l_z and l^*_z statistics. (The results for ζ_1, ζ^*_1, ζ_2, and ζ^*_2 are similar and are available upon request from the first author.) Quintiles were formed by separating examinees based on θ, such that low-ability examinees were placed in Quintile 1, high-ability examinees were placed in Quintile 5, and the examinees in between were sorted accordingly. Notably, the Type I error rates of the newly proposed statistics are similar across ability levels.

3.2.3. Power Table 1 displays the power of l_z and l^*_z at the $\alpha = .01$ significance level. (The results for ζ_1, ζ^*_1, ζ_2, and ζ^*_2 are similar and are available upon request.) Each row corresponds to a different combination of aberrance and test length, and each column corresponds to a different combination of person-fit statistic and skewness correction. As expected, power increases as test length increases and as the proportion of contaminated items increases. Furthermore, across all

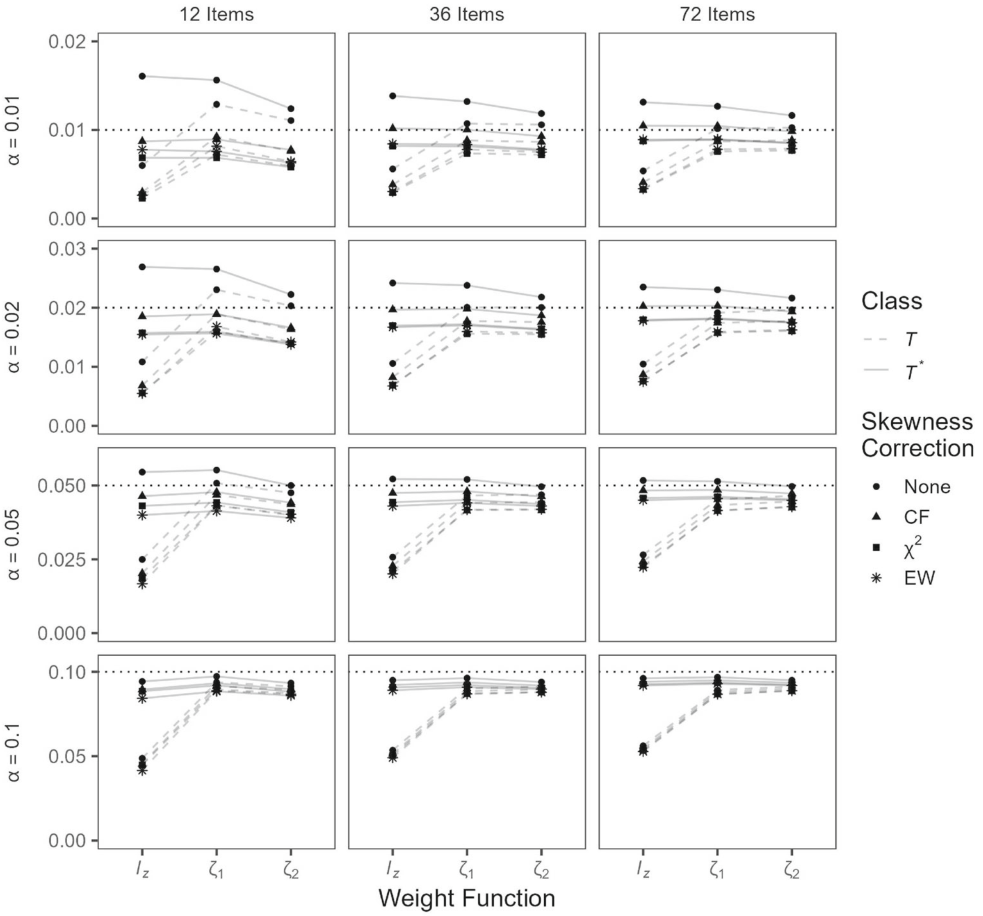

FIGURE 4.
Type I error rates of the person-fit statistics. *CF* Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

conditions, the l_z^* statistic with no skewness correction is the most powerful, followed closely by the l_z^* statistic corrected using the Cornish–Fisher expansion, the l_z^* statistic corrected using the Edgeworth expansion, and then the l_z^* statistic corrected using the χ^2 approximation. Thus, it appears that the new statistics can be used without a significant loss in power. Similar results are shown in Table 2 at the $\alpha = .05$ significance level.

4. Real Data Example

4.1. Data and Analysis

The data in this example originate from a single form of a licensure examination. The data have been studied in the context of person-fit assessment by researchers such as Sinharay (2016b), and in the context of preknowledge detection in several chapters of Cizek & Wollack (2017). Item scores are available for 1644 examinees on 170 scored items. Following a statistical analysis and careful investigative process, the testing program flagged 61 items as being compromised, and 48 examinees as likely having engaged in fraudulent behavior. The 48 flagged examinees can be

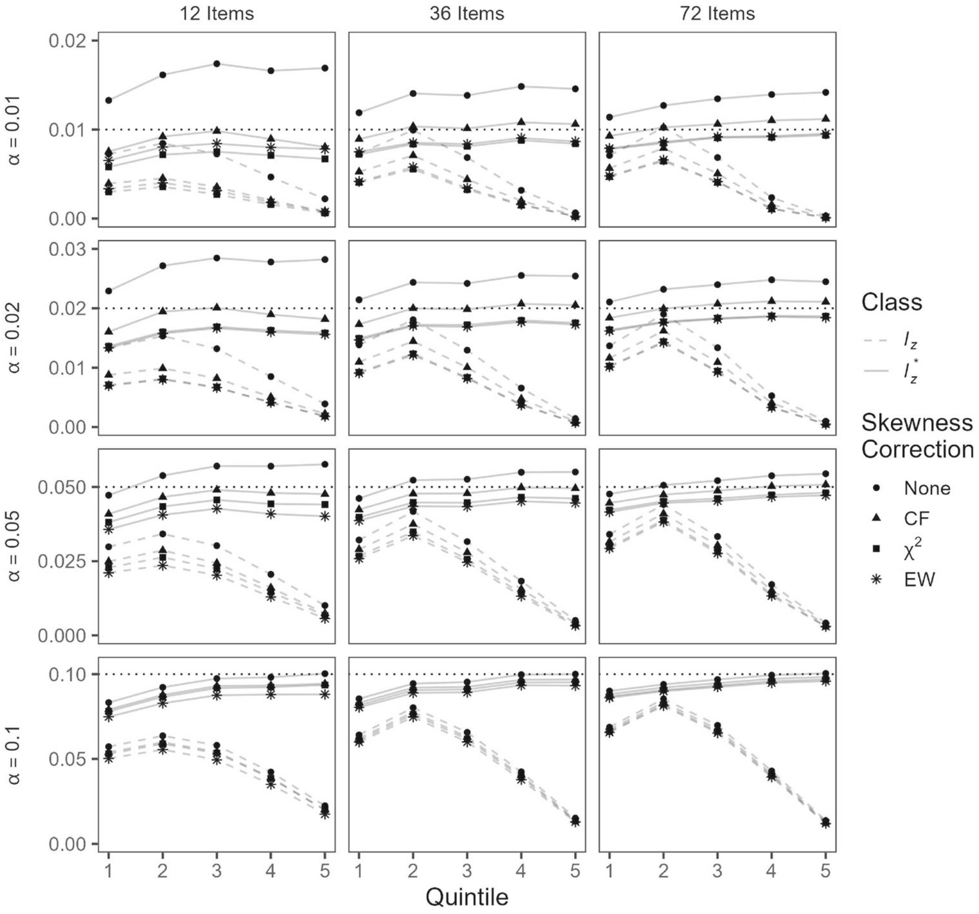

FIGURE 5.
Type I error rates by quintile of l_z and l_z^*. *CF* Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

considered truly aberrant for purposes of the present analysis. However, it is important to note that other types of aberrance may be present among some of the non-flagged examinees, as well.

The Rasch model parameter estimates provided by the testing program were treated as the true item parameters. Then, using the WL estimates of ability, each score pattern was analyzed 24 times: once for each combination of two classes of person-fit statistics (T, T^*), three weight functions (l_z, ζ_1, ζ_2), and four skewness corrections (none, Cornish–Fisher expansion, χ^2 approximation, Edgeworth expansion).

4.2. Results

Tables 3 and 4 display the proportions of examinees classified as aberrant and the agreement rates for l_z and l_z^* at the $\alpha = .01$ and $\alpha = .05$ significance levels, respectively. (The results for ζ_1, ζ_1^*, ζ_2, and ζ_2^* are similar and are available upon request.) In each table, the proportions of examinees classified as aberrant are displayed in bold text along the diagonal, and the agreement rates are displayed in non-bold text in the off-diagonal. Agreement rate is defined as the proportion of times two statistics make the same classification decision (aberrant or non-aberrant).

TABLE 1.
Power of l_z and l_z^* ($\alpha = .01$).

		l_z				l_z^*			
Aberrance	Test length	None	CF	χ^2	EW	None	CF	χ^2	EW
Lack of motivation on $\frac{1}{6}$	12	.106	.063	.056	.063	.190	.142	.129	.135
	36	.383	.341	.314	.318	.476	.441	.417	.421
	72	.644	.625	.608	.609	.704	.687	.672	.674
Lack of motivation on $\frac{1}{3}$	12	.177	.133	.120	.126	.250	.201	.183	.190
	36	.478	.451	.432	.435	.567	.538	.518	.520
	72	.653	.638	.626	.627	.748	.735	.726	.727
Item disclosure on $\frac{1}{6}$	12	.041	.022	.017	.019	.093	.053	.043	.048
	36	.158	.131	.114	.117	.281	.241	.213	.217
	72	.355	.325	.303	.305	.553	.520	.494	.497
Item disclosure on $\frac{1}{3}$	12	.108	.069	.058	.064	.229	.156	.135	.145
	36	.317	.279	.259	.264	.566	.523	.494	.498
	72	.494	.464	.445	.447	.795	.776	.760	.762

CF Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

TABLE 2.
Power of l_z and l_z^* ($\alpha = .05$).

		l_z				l_z^*			
Aberrance	Test length	None	CF	χ^2	EW	None	CF	χ^2	EW
Lack of motivation on $\frac{1}{6}$	12	.233	.208	.198	.187	.341	.318	.309	.299
	36	.575	.560	.550	.544	.649	.635	.625	.622
	72	.772	.766	.761	.760	.823	.818	.814	.813
Lack of motivation on $\frac{1}{3}$	12	.314	.294	.284	.274	.401	.379	.370	.362
	36	.613	.602	.594	.592	.702	.692	.687	.683
	72	.762	.756	.752	.751	.844	.840	.837	.836
Item disclosure on $\frac{1}{6}$	12	.128	.109	.102	.093	.237	.211	.199	.187
	36	.325	.308	.295	.290	.509	.488	.476	.470
	72	.549	.534	.524	.520	.759	.749	.740	.738
Item disclosure on $\frac{1}{3}$	12	.239	.212	.203	.190	.436	.401	.386	.372
	36	.476	.456	.445	.437	.743	.731	.721	.716
	72	.647	.635	.626	.622	.892	.888	.885	.884

CF Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

Across all person-fit statistics and skewness corrections, the proportions of flagged examinees classified as aberrant are much larger than the proportions of non-flagged examinees classified as aberrant. This result provides favorable evidence regarding the performance of the statistics. In addition, the proportions of non-flagged examinees classified as aberrant consistently exceed the significance levels. This result is interesting, as it implies that aberrance is present among some of the non-flagged examinees, as well. For example, some of the non-flagged examinees may have engaged in fraudulent behavior, but were mistakenly not flagged by the testing program. It is also possible that some of the non-flagged examinees had engaged in a different type of aberrant behavior altogether.

Consistent with the simulation results, the l_z^* statistic with no skewness correction classified the most examinees as aberrant, followed by the l_z^* statistic corrected using the Cornish–Fisher

TABLE 3.
Proportions of examinees classified as aberrant (diagonal) and agreement rates (off-diagonal) ($\alpha = .01$).

			l_z				l_z^*			
Group	Class	Correction	None	CF	χ^2	EW	None	CF	χ^2	EW
Non-Flagged	l_z	None	**.088**							
		CF	.994	**.082**						
		χ^2	.990	.996	**.078**					
		EW	.991	.996	.999	**.078**				
	l_z^*	None	.968	.962	.958	.959	**.120**			
		CF	.973	.967	.963	.964	.995	**.115**		
		χ^2	.976	.972	.971	.971	.987	.992	**.107**	
		EW	.975	.972	.970	.971	.988	.993	.999	**.108**
Flagged	l_z	None	**.167**							
		CF	1.000	**.167**						
		χ^2	1.000	1.000	**.167**					
		EW	1.000	1.000	1.000	**.167**				
	l_z^*	None	.896	.896	.896	.896	**.271**			
		CF	.896	.896	.896	.896	1.000	**.271**		
		χ^2	.896	.896	.896	.896	1.000	1.000	**.271**	
		EW	.896	.896	.896	.896	1.000	1.000	1.000	**.271**

CF Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

TABLE 4.
Proportions of examinees classified as aberrant (diagonal) and agreement rates (off-diagonal) ($\alpha = .05$).

			l_z				l_z^*			
Group	Class	Correction	None	CF	χ^2	EW	None	CF	χ^2	EW
Non-Flagged	l_z	None	**.147**							
		CF	.998	**.145**						
		χ^2	.997	.999	**.145**					
		EW	.997	.999	.999	**.144**				
	l_z^*	None	.944	.942	.942	.941	**.203**			
		CF	.950	.948	.947	.947	.994	**.197**		
		χ^2	.953	.951	.951	.950	.991	.997	**.194**	
		EW	.954	.952	.951	.951	.991	.996	.999	**.194**
Flagged	l_z	None	**.250**							
		CF	1.000	**.250**						
		χ^2	1.000	1.000	**.250**					
		EW	.979	.979	.979	**.229**				
	l_z^*	None	.896	.896	.896	.875	**.354**			
		CF	.896	.896	.896	.875	1.000	**.354**		
		χ^2	.896	.896	.896	.875	1.000	1.000	**.354**	
		EW	.896	.896	.896	.875	1.000	1.000	1.000	**.354**

CF Cornish–Fisher expansion, χ^2 χ^2 approximation, *EW* Edgeworth expansion.

expansion. Notably, all four variants of the l_z^* statistic classified the same sets of flagged examinees as aberrant—however, the skewness-corrected statistics classified fewer non-flagged examinees as aberrant. This result shows that the new statistics may lead to noticeable differences in tests having as many as 170 items.

5. Discussion

Many popular person-fit statistics—including l_z, ζ_1, and ζ_2—belong to the class of standardized person-fit statistics, T, and are assumed to have a standard normal null distribution. However, in practice, this assumption is incorrect since T is computed using (a) an estimated ability parameter and (b) a finite number of items. In this paper, we proposed three new corrections for T that simultaneously account for the use of an estimated ability parameter and the use of a finite number of items. The new corrections are efficient in that they only require the analysis of the original data set and do not require the simulation or analysis of any additional data sets (as is the case for resampling-based methods). Detailed simulations further revealed that the new corrections are able to control the Type I error rate while also maintaining reasonable levels of power. They therefore outperform the existing corrections for T that were suggested by Bedrick (1997), Molenaar & Hoijtink (1990), and Snijders (2001).

Based on the results of the simulation study, we created the following set of guidelines for users to follow while selecting an appropriate person-fit statistic:

- When $\alpha \geq .10$, it is recommended that users apply the existing T^* statistic of Snijders (2001).
- When $\alpha < .10$, it is recommended that users apply the newly proposed T^*_{CF} statistic.

Note that the recommended statistics are those that were shown to display the largest power while still controlling the Type I error rate.

We would also like to remind readers that person-fit statistics, by definition, are most appropriate when the goal is to detect general misfit at the person level. If the goal is to detect a specific type of misfit, such as item preknowledge or test speededness, or if the goal is to detect misfit at the person-by-item level, then alternative methods may be more suitable.

There are several limitations to this work, providing many opportunities for future research. First, it is possible to study shorter tests, to explore the boundaries of the proposed methods to see if there is a point at which they fail or break down. It is also possible to study additional simulation conditions. For example, we simulated lack of motivation on the easiest items and item disclosure on the most difficult items, thereby considering extreme conditions in which ability estimates are severely impacted. However, in practice, such behaviors could happen on more than just the easiest and most difficult items. Therefore, future researchers could simulate less extreme conditions to compare the statistics under a more realistic setting. In our simulations, we also assumed that the item parameters were known. However, researchers such as Cheng & Yuan (2010) have shown that the error associated with item parameter estimation affects the distribution of the person parameter estimates. It would be interesting to study the extent to which this error affects the distributions of the person-fit statistics, as well.

Second, the efficient corrections that are described in this paper could be compared to the resampling-based methods that are described in Sinharay (2016a). The methods should be compared in terms of the false-positive rate, true-positive rate, and computation time. Third, the new corrections could be applied to other standardized person-fit statistics, such as the standardized infit and outfit statistics (Magis et al. 2014). Fourth, the new corrections could be applied using other ability estimates, such as the biweight estimate and the Huber estimate. Sinharay (2016d) found that the use of both estimates with T^* produces inflated Type I error rates, suggesting that the new corrections may be particularly useful in these settings.

Finally, in this study, we developed corrections for the class of standardized person-fit statistics within a very narrow context: non-adaptive, unidimensional tests with only dichotomous items. Standardized person-fit statistics have been applied in other contexts, including adaptive tests (e.g., Nering 1997; van Krimpen-Stoop & Meijer 1999), multidimensional tests with simple structure (e.g., Albers et al. 2016; Hong et al. 2021), tests with polytomous items (e.g., Gorney & Wollack 2023; Hong et al. 2021; Sinharay 2016c; van Krimpen-Stoop & Meijer 2002; von Davier & Molenaar 2003), and tests with response times (e.g., Gorney et al. 2024). Standardized person-fit statistics have also been applied in cognitive diagnosis modeling (e.g., Santos et al. 2020). In each of these contexts, the assumption of the standard normal null distribution has been shown to be inappropriate when realistic testing conditions are simulated, suggesting that corrections may be beneficial.

Funding This work was completed while the first author was an Educational Testing Service (ETS) Harold Gulliksen Psychometric Research Fellow.

Conflict of interest The authors declare that they have no conflict of interest.

Data Availability The data that support the findings of this study are available from Dr. James Wollack upon reasonable request.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix A

R Code to Compute the Person-Fit Statistics

```
# INPUT
# `x` is an (N x n) matrix of item scores. The rows correspond to persons and
#   the columns correspond to items.
# `psi` is an (n x 3) matrix of item parameters. The columns correspond to the
#   discrimination, difficulty, and pseudo-guessing parameters, respectively.
# `theta` is a vector of person parameter estimates.
# `est` is the person parameter estimation method. Options are "WL" for weighted
#   likelihood estimation, "ML" for maximum likelihood estimation, and "MAP" for
#   maximum a posteriori estimation with a standard normal prior.
# `weight` is the weight function. Options are "lz", "zeta1", and "zeta2".
#
# OUTPUT
# A matrix of eight standardized person-fit statistics.

personfit <- function(x, psi, theta, est = "ML", weight = "lz") {

  # Setup
  N <- nrow(x)   # number of examinees
  n <- ncol(x)   # number of items
  stat <-
  matrix(nrow = N, ncol = 8,
         dimnames = list(
           person = 1:N,
           method = c("NO", "CF", "CS", "EW", "TS", "TSCF", "TSCS", "TSEW")
         ))

# IRT probabilities (Eq. 1)
p <- t(psi[, 3] + (1 - psi[, 3]) /
         (1 + exp(psi[, 1] * outer(psi[, 2], theta, "-"))))
q <- 1 - p

# Weight functions (Eqs. 4 and 5)
if (weight == "lz") {
  w <- log(p / q)
} else if (weight == "zeta1") {
  w <- matrix(mean(p) - colMeans(p), nrow = N, ncol = n, byrow = TRUE)
} else if (weight == "zeta2") {
  w <- rowMeans(p) - p
}

# (Eq. 3)
W <- rowSums((x - p) * w)

# Mean, standard deviation, and skewness (Eqs. 6, 7, and 8)
mu <- 0
sigma <- sqrt(rowSums(p * q * w^2))
gamma <- rowSums(p * q * (q - p) * w^3) / sigma^3

# No correction (Eq. 10)
stat[, "NO"] <- NO <- (W - mu) / sigma

# Cornish-Fisher expansion (Eq. 20)
stat[, "CF"] <- NO - gamma * (NO^2 - 1) / 12

# Chi-squared approximation (Eq. 22)
nu <- 8 / gamma^2
b <- sqrt(sigma^2 / (2 * nu))
a <- b * nu
if (weight == "lz") {
  p_CS <- pchisq(abs(W - mu - a) / b, df = nu, lower.tail = FALSE)
  stat[, "CS"] <- qnorm(p_CS)
} else {
  p_CS <- pchisq(abs(W - mu + a) / b, df = nu, lower.tail = FALSE)
  stat[, "CS"] <- qnorm(1 - p_CS)
}

# Edgeworth expansion (Eq. 23)
if (weight == "lz") {
  p_EW <- pnorm(NO) - dnorm(NO) * gamma * (NO^2 - 1) / 6
  p_EW <- ifelse(p_EW <= 0 | p_EW >= 1, pnorm(NO, lower.tail = TRUE), p_EW)
  stat[, "EW"] <- qnorm(p_EW)
} else {
  p_EW <- 1 - (pnorm(NO) - dnorm(NO) * gamma * (NO^2 - 1) / 6)
  p_EW <- ifelse(p_EW <= 0 | p_EW >= 1, pnorm(NO, lower.tail = FALSE), p_EW)
  stat[, "EW"] <- qnorm(1 - p_EW)
}
# Various quantities
```

```
  e <- exp(-psi[, 1] * outer(psi[, 2], theta, "-"))
  p1 <- t((1 - psi[, 3]) * psi[, 1] * e / (1 + e)^2)
  p2 <- t((1 - psi[, 3]) * psi[, 1]^2 * e * (1 - e) / (1 + e)^3)
  r <- p1 / (p * q)
  c <- rowSums(p1 * w) / rowSums(p1 * r)
  r0 <- if (est == "ML") 0 else if (est == "MAP") -theta else
    rowSums(r * p2) / (2 * rowSums(r * p1))

  # Modified weight function (Eq. 17)
  w_tilde <- w - c * r

  # Mean, standard deviation, and skewness (Eqs. 14, 15, and 24)
  mu_tilde <- -c * r0
  sigma_tilde <- sqrt(rowSums(p * q * w_tilde^2))
  gamma_tilde <- rowSums(p * q * (q - p) * w_tilde^3) / sigma_tilde^3

  # Taylor series expansion (Eq. 18)
  stat[, "TS"] <- TS <- (W - mu_tilde) / sigma_tilde

  # Taylor series expansion and Cornish-Fisher expansion (Eq. 25)
  stat[, "TSCF"] <- TS - gamma_tilde * (TS^2 - 1) / 12

  # Taylor series expansion and chi-squared approximation (Eq. 27)
  nu_tilde <- 8 / gamma_tilde^2
  b_tilde <- sqrt(sigma_tilde^2 / (2 * nu_tilde))
  a_tilde <- b_tilde * nu_tilde
  if (weight == "lz") {
    p_TSCS <- pchisq(abs(W - mu_tilde - a_tilde) / b_tilde,
                     df = nu_tilde, lower.tail = FALSE)
    stat[, "TSCS"] <- qnorm(p_TSCS)
  } else {
    p_TSCS <- pchisq(abs(W - mu_tilde + a_tilde) / b_tilde,
                     df = nu_tilde, lower.tail = FALSE)
    stat[, "TSCS"] <- qnorm(1 - p_TSCS)
  }

  # Taylor series expansion and Edgeworth expansion (Eq. 28)
  if (weight == "lz") {
    p_TSEW <- pnorm(TS) - dnorm(TS) * gamma_tilde * (TS^2 - 1) / 6
    p_TSEW <- ifelse(p_TSEW <= 0 | p_TSEW >= 1,
                     pnorm(TS, lower.tail = TRUE), p_TSEW)
    stat[, "TSEW"] <- qnorm(p_TSEW)
  } else {
    p_TSEW <- 1 - (pnorm(TS) - dnorm(TS) * gamma_tilde * (TS^2 - 1) / 6)
    p_TSEW <- ifelse(p_TSEW <= 0 | p_TSEW >= 1,
                     pnorm(TS, lower.tail = FALSE), p_TSEW)
    stat[, "TSEW"] <- qnorm(1 - p_TSEW)
  }

  # Output
  return(stat)
}
```

Appendix B

Ability Estimates

See Figs. 6 and 7.

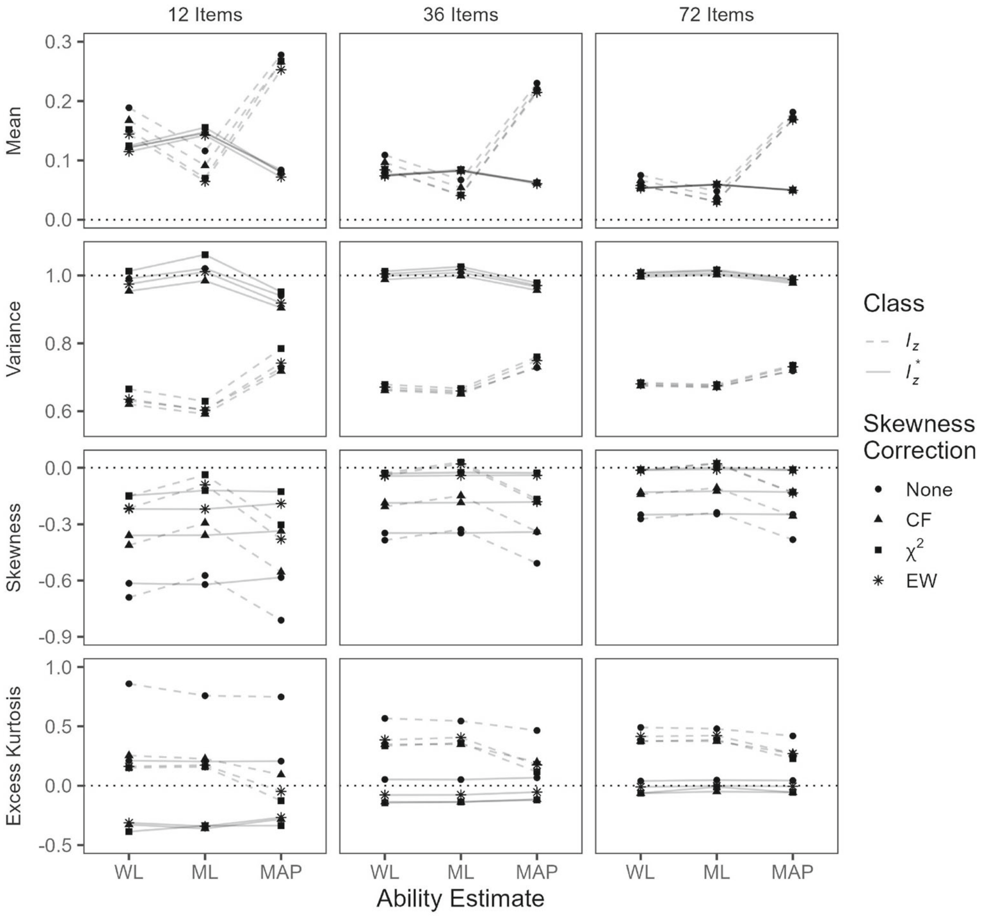

FIGURE 6.
Descriptive statistics of the null distributions of l_z and l_z^*.

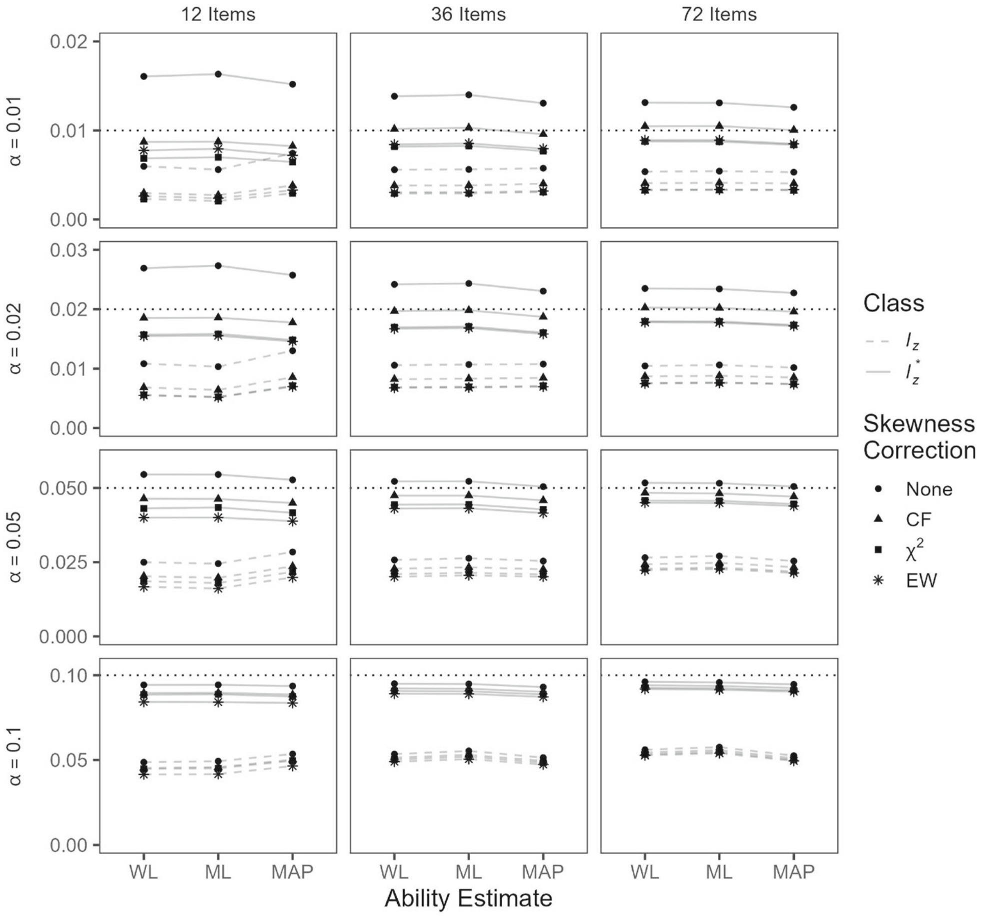

FIGURE 7.
Type I error rates of l_z and l_z^*.

References

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

Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika, 62*(2), 191–199. https://doi.org/10.1007/BF02295274

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75*(2), 280–291. https://doi.org/10.1007/s11336-009-9144-x

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge. https://doi.org/10.4324/9781315743097

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. https://doi.org/10.1111/j.1745-3984.2008.00058.x

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

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*(3), 217–233. https://doi.org/10.1177/0146621603027003003

Gorney, K., Sinharay, S., & Liu, X. (2024). Using item scores and response times in person-fit assessment. *British Journal of Mathematical and Statistical Psychology, 77*(1), 151–168. https://doi.org/10.1111/bmsp.12320

Gorney, K., & Wollack, J. A. (2023). Using item scores and distractors in person-fit assessment. *Journal of Educational Measurement, 60*(1), 3–27. https://doi.org/10.1111/jedm.12345

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. https://doi.org/10.1007/s11336-021-09756-3

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215–231. https://doi.org/10.1177/01466216970213002

Magis, D., Béland, S., & Raîche, G. (2014). Snijders's correction of the infit and outfit indices with estimated ability level: An analysis with the Rasch model. *Journal of Applied Measurement, 15*(1), 82–93.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*(1), 57–81. https://doi.org/10.3102/1076998610396894

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75–106. https://doi.org/10.1007/BF02294745

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*(2), 115–127. https://doi.org/10.1177/01466216970212002

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16*(4), 345–352. https://doi.org/10.1177/014662169201600405

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*(3), 213–229. https://doi.org/10.1177/014662169501900301

Santos, K. C. P., de la Torre, J., & von Davier, M. (2020). Adjusting person fit index for skewness in cognitive diagnosis modeling. *Journal of Classification, 37*(2), 399–420. https://doi.org/10.1007/s00357-019-09325-5

Sinharay, S. (2016a). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement, 53*(1), 63–85. https://doi.org/10.1111/jedm.12101

Sinharay, S. (2016b). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement, 40*(6), 418–433. https://doi.org/10.1177/0146621616649963

Sinharay, S. (2016c). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81*(4), 992–1013. https://doi.org/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*(2), 175–193. https://doi.org/10.1111/bmsp.12067

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342. https://doi.org/10.1007/BF02294437

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95–110. https://doi.org/10.1007/BF02294208

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*(4), 327–345. https://doi.org/10.1177/01466219922031446

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*(2), 164–180. https://doi.org/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*(2), 213–228. https://doi.org/10.1007/BF02294798

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Manuscript Received: 11 MAR 2023
Published Online Date: 1 APR 2024