Asymptotically Normally Distributed Person Fit Indices for Detecting Spuriously High Scores on Difficult Items

Applied Psychological Measurement

2018, Vol. 42(5) 343–358

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617730391

journals.sagepub.com/home/apm

Yan Xia¹ and Yi Zheng¹

Abstract

Snijders developed a family of person fit indices that asymptotically follow the standard normal distribution, when the ability parameter is estimated. So far, I_z^* , U^* , W^* , $ECI2_z^*$, and $ECI4_z^*$ from this family have been proposed in previous literature. One common property shared by I_z^* , U^* , and W^* (also $ECI2_z^*$ and $ECI4_z^*$ in some specific conditions) is that they employ symmetric weight functions and thus identify spuriously high scores on both easy and difficult items in the same manner. However, when the purpose is to detect only the spuriously high scores on difficult items, such as cheating, guessing, and having item preknowledge, using symmetric weight functions may jeopardize the detection rates of the target aberrant response patterns. By specifying two types of asymmetric weight functions, this study proposes $SHa(\lambda)^*$ ($\lambda = 1/2$ or 1) and $SHb(\beta)^*$ ($\beta = 2$ or 3) based on Snijders's framework to specifically detect spuriously high scores on difficult items. Two simulation studies were carried out to investigate the Type I error rates and empirical power of $SHa(\lambda)^*$ and $SHb(\beta)^*$, compared with I_z^* , U^* , W^* , $ECI2_z^*$, and $ECI4_z^*$. The empirical results demonstrated satisfactory performance of the proposed indices. Recommendations were also made on the choice of different person fit indices based on specific purposes.

Keywords

item response theory, person fit statistic, asymptotically standard normal distribution

When item response theory (IRT) is used to score examinees, latent trait scores can be misleading due to aberrant response behaviors such as cheating in educational assessments or answering under the influence of social desirability in psychological assessments. A number of person fit indices have been developed to detect such aberrant response behaviors. For example, one of the most commonly known indices, I_z (Drasgow, Levine, & Williams, 1985), is expressed as the standardized log-likelihood of a response pattern based on a certain IRT model. Molenaar and Hoijtink (1990, 1996) showed that I_z asymptotically follows the standard normal distribution when the true value of the ability parameter is used. In practice, however, the true ability value

¹Arizona State University, Tempe, USA

Corresponding Author:

Yan Xia Postdoctoral Researcher T. Denny Sanford School of Social and Family Dynamics Arizona State University
951 Cady Mall, Tempe, AZ 85287.

Email: yxia@asu.edu

is unavailable, and substituting it with its estimate will cause I_z to deviate from the standard normal distribution, with a variance smaller than 1 (Drasgow et al., 1985; Nering, 1995; Reise, 1995; Snijders, 2001). Therefore, when the ability parameter is estimated, aberrant response detection using the critical values from the standard normal distribution can be problematic, in particular, having less power (Meijer & Sijtsma, 2001).

To address the inaccuracy caused by using the estimated ability parameter instead of the true value, Snijders (2001) proposed a solution by adjusting the mean and variance of I_z . The resulting person fit index, I_z^* , asymptotically¹ follows the standard normal distribution when the estimated ability parameter is used. Through simulation studies, Snijders found that I_z^* has more accurate Type I error rates compared with I_z when the critical values are chosen from the standard normal distribution. I_z^* has received growing attention. Recently, Magis, Raïche, and Béland (2012) provided a didactic presentation of I_z^* and made their R codes available. The R package “PerFit” also provides the calculation of I_z^* (Tendeiro, Meijer, & Niessen, 2015).

In addition to I_z^* , Snijders’s (2001) framework can be used to formulate a family of person fit indices, all expressed as weighted sum of squared residuals. Meijer and Sijtsma (2001) and Snijders (2001) mentioned that several existing person fit indices can be transformed into what follows the asymptotic standard normal distribution. Magis, Béland, and Raïche (2014) generalized U and W (Wright & Stone, 1979; see also Wright & Masters, 1982) to U^* and W^* , respectively, in the Rasch model, and found that the latter two had well-controlled Type I error rates and improved power compared with the former. Sinharay (2016a) generalized two standardized extended caution indices, $ECI2_z$ and $ECI4_z$ (Tatsuoka, 1984; Tatsuoka & Tatsuoka, 1982), to $ECI2_z^*$ and $ECI4_z^*$. Sinharay’s (2016a) simulation study showed that $ECI2_z^*$ and $ECI4_z^*$ led to more accurate Type I error rates and improved power compared with the uncorrected $ECI2_z$ and $ECI4_z$.

I_z^* , U^* , and W^* share a common property that they employ symmetric weight functions and thus have equal sensitivity to spuriously high scores on difficult items and spuriously low scores on easy items. For example, suppose examinee A has a 5% chance of responding correctly to a difficult item, and examinee B has a 95% chance of responding correctly to an easy item. In addition, suppose that examinee A answers the difficult item correctly (e.g., by having item pre-knowledge) and examinee B fails the easy item (e.g., by careless responding). Then the weights attached to the above two aberrant responses are equal for all the three above-mentioned indices. Therefore, for I_z^* , U^* , and W^* , spuriously high scores and spuriously low scores are both monitored in the same manner. As discussed later, the weight functions of $ECI2_z^*$ and $ECI4_z^*$ can also be symmetric in some specific conditions, depending on the ability and item parameters.

For many assessments, however, the above two aberrant cases are not equally concerning. For example, many educational assessments aim to select individuals with high ability levels. Thus, spuriously high scores are a more serious threat to the validity of the test. There also tend to be more attempts to achieve spuriously high scores than spuriously low scores. Similarly, some psychological tests aim to screen out people with the lowest latent trait levels, and thus, spuriously low scores more likely lead to incorrect conclusions and should be given more attention. However, because I_z^* , U^* , and W^* (as well as $ECI2_z^*$ and $ECI4_z^*$ given specific person and item parameters) split their power equally between spuriously high scores and spuriously low scores, consequently, they lose power when only one side of the scale should be the focus of attention.

The main purpose of the current study is to propose, based on Snijders’s (2001) framework, two new person fit indices that aim to detect spuriously high scores with greater power. These indices are designed for tests where the authors wish to more effectively detect spuriously high scores and have less concern over spuriously low scores. Also note that the proposed indices

can be easily adapted to focus on detecting spuriously low scores instead. This will be elaborated in the “Conclusions and Discussion” section.

To denote the two types of indices the authors are proposing, $\text{SHa}(\lambda)^*$ (Spuriously High—Type a) and $\text{SHb}(\beta)^*$ (Spuriously High—Type b), where λ and β are two constants that are explained in the “Introducing $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ ” section. $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ both employ asymmetric weight functions with greater weights to the responses on difficult items. Therefore, $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ focus on detecting spuriously high scores on difficult items and can potentially provide improved power thereof. In addition, because $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ are derived following Snijders’s (2001) framework, they both follow standard normal distributions asymptotically under the null conditions and are expected to result in well-controlled Type I error rates given long enough tests.

The second purpose of this study is to compare the performance of $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ with the existing person fit indices that fall within Snijders’s (2001) framework, including I_z^* , U^* , W^* , ECI2_z^* , and ECI4_z^* . Magis et al. (2014) simulated data from tests with 40 and 80 items using Rasch models and compared U^* , W^* , and their uncorrected versions. They found that W^* was superior to U^* with respect to the accuracy of Type I error rates and the magnitude of power in detecting both cheating (spuriously high scores) and inattention (spuriously low scores). Sinharay (2016a) compared ECI2_z^* , ECI4_z^* , their uncorrected versions, and I_z^* under three-parameter logistic models and various test length conditions (20, 40, and 60 items). Sinharay (2016a) simulated data with and guessing on easy items (spuriously low scores), and they found that (a) ECI2_z^* led to inflated Type I error rates but was the most sensitive in detecting spuriously high scores, (b) ECI4_z^* had the most accurate Type I error rates but lowest power to detect spuriously low scores, and (c) I_z^* had the most inaccurate Type I error rates but the largest power in detecting spuriously low scores on the 10% easiest items. Because these corrected indices, I_z^* , U^* , W^* , ECI2_z^* , and ECI4_z^* , were all found superior compared with their uncorrected versions, these corrected versions were chosen to be compared with $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ in this study.

The following section first describes the theoretical background following Snijders’s (2001) framework. Then, the next section provides an integrated formulation that expresses I_z^* , U^* , W^* , ECI2_z^* , and ECI4_z^* all as weighted sums of squared residuals, characterized by their different weight functions. $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ are then presented, followed by two simulation studies: Study I investigates the empirical Type I error rates of the above-mentioned indices when no aberrant response exists, and Study II examines the power of them when spuriously high scores are present. The closing section provides a summary of the findings and conclusions for the current research.

Theoretical Background

Let N be the number of items, X_n be the binary score for the n th item where 0 indicates an incorrect answer and 1 indicates a correct answer, θ be the true value of the ability parameter, $P_n(\theta)$ be the probability of a correct answer for the n th item given θ , and $Q_n(\theta)$ be the probability of an incorrect answer for the n th item given θ . For the one-, two-, and three-parameter logistic models (1PLM, 2PLM, and 3PLM, respectively), $P_n(\theta)$ can be expressed as $P_n(\theta) = c_n + (1 - c_n)(1 + \exp^{-1}[a_n(\theta - b_n)])^{-1}$. For the 3PLM, c_n is the probability of answering 1 with a negative infinite θ value for the n th item. When $c_n = 0$ for all n , the model reduces to the 2PLM. When $c_n = 0$ and $a_{n1} = a_{n2}$ for every pair of distinct $n1$ and $n2$, the model reduces further to the 1PLM.

As summarized by Meijer and Sijtsma (2001), one family of parametric person fit indices can be expressed as the weighted sum of squared residuals:

$$V(\theta) = \sum_{n=1}^N [X_n - P_n(\theta)]^2 v_n(\theta), \quad (1)$$

where $v_n(\theta)$ is a suitable weight function. Because X_n is a binary variable, $X_n^2 = X_n$. $V(\theta)$ can therefore be reexpressed as

$$V(\theta) = \sum_{n=1}^N [X_n - P_n(\theta)][1 - 2P_n(\theta)]v_n(\theta) + \sum_{n=1}^N P_n(\theta)Q_n(\theta)v_n(\theta). \quad (2)$$

The second term on the right side of Equation 2, $\sum_{n=1}^N P_n(\theta)Q_n(\theta)v_n(\theta)$, is a constant given θ , N , and item parameters, and it will not contribute to the detection of aberrant response patterns.

Let

$$\omega_n(\theta) = [1 - 2P_n(\theta)]v_n(\theta). \quad (3)$$

Using Equation 3 and ignoring the second term on the right side of Equation 2, $V(\theta)$ can be reexpressed in the following centered form:

$$W(\theta) = \sum_{n=1}^N [X_n - P_n(\theta)]\omega_n(\theta). \quad (4)$$

The connection between Equations 1 and 4 is expressed by Equation 3. Once a person fit statistic with its weight function $v_n(\theta)$ in Form 1 is known, it can be reexpressed in the centered form in Equation 4 using Equation 3.

Based on Equation 4, $E[W(\theta)] = 0$ and $Var[W(\theta)] = N\sigma_N^2(\theta)$, with $\sigma_N^2(\theta) = \sum_{n=1}^N \omega_n(\theta)^2 P_n(\theta)Q_n(\theta)/N$. Then, the person fit statistic following the asymptotic standard normal distribution can be expressed as $W(\theta)/[Var[W(\theta)]]^{1/2}$, when the true value of θ is used (Snijders, 2001).

When the ability parameter is estimated, a correction should be made to keep the distribution asymptotically standard normal. Below, a general description of the procedure for correcting the estimated ability parameter $\hat{\theta}$, without detailed explanation, is shown. Readers who are interested in the details can refer Snijders (2001) and the didactic presentation by Magis et al. (2012).

In practice, θ is estimated and replaced by $\hat{\theta}$, satisfying $r_0(\hat{\theta}) + \sum_{n=1}^N [X_n - P_n(\hat{\theta})]r_n(\hat{\theta}) = 0$. Let $P_n'(\hat{\theta})$ and $P_n''(\hat{\theta})$ be the first and second derivatives, respectively, of $P_n(\hat{\theta})$ with respect to $\hat{\theta}$; $f(\hat{\theta})$ be a prior density function (e.g., a normal distribution density function); and $I(\hat{\theta})$ and $J(\hat{\theta})$ be $I(\hat{\theta}) = \sum_{n=1}^N P_n'(\hat{\theta})^2 / P_n(\hat{\theta})Q_n(\hat{\theta})$ and $J(\hat{\theta}) = \sum_{n=1}^N P_n'(\hat{\theta})P_n''(\hat{\theta}) / P_n(\hat{\theta})Q_n(\hat{\theta})$, respectively. $r_0(\hat{\theta})$ is 0, $d \log f(\hat{\theta}) / d\hat{\theta}$, or $J(\hat{\theta}) / 2I(\hat{\theta})$ for the maximum likelihood estimator (MLE), maximum a posteriori (MAP) estimator, or the weighted likelihood (WLE) estimator, respectively. $r_n(\hat{\theta})$ is 1, a_n , or $P_n'(\hat{\theta}) / [P_n(\hat{\theta})Q_n(\hat{\theta})]$ for the 1PLM, 2PLM, or 3PLM, respectively.

When $\hat{\theta}$ is used instead of θ , Snijders (2001) showed that $W(\hat{\theta}) / [Var[W(\hat{\theta})]]^{1/2}$ deviates from the standard normal distribution. However, by adjusting the mean and variance of $W(\hat{\theta})$ using

$$W^*(\hat{\theta}) = \frac{W(\hat{\theta}) + c_N(\hat{\theta})r_0(\hat{\theta})}{N^{1/2}\tau_N(\hat{\theta})}, \quad (5)$$

$W^*(\hat{\theta})$ follows asymptotically the standard normal distribution. $\omega_n(\theta)$ is constraint such that it is twice differentiable. In addition, $\omega_n(\theta)$ and its first and second derivatives should be uniformly

bounded in a neighborhood of θ . In Equation 5, $c_N(\hat{\theta}) = \sum_{n=1}^N P'_n(\hat{\theta})\omega_n(\hat{\theta}) / \sum_{n=1}^N P'_n(\hat{\theta})r_n(\hat{\theta})$ and $\tau_N^2(\hat{\theta}) = \sum_{n=1}^N [\omega_n(\hat{\theta}) - c_N(\hat{\theta})r_n(\hat{\theta})]^2 \cdot P_n(\hat{\theta})Q_n(\hat{\theta}) / N$.

As long as a person fit statistic can be expressed in Form 1, it can be reexpressed in the centered Form 4, and then its asymptotic version can be derived using Equation 5. The math operations from Equations 1 to 5 are linear transformations, without changing the direction of the indices (i.e., a larger value of $V(\hat{\theta})$ corresponds to a larger value of $W^*(\hat{\theta})$). Conceptual explanations for $W^*(\hat{\theta})$ can therefore be based on Form 1. Furthermore, in Form 1, multiplying a constant with the weight function $v_n(\hat{\theta})$ or adding a constant to $V(\hat{\theta})$ will not influence the values of $W^*(\hat{\theta})$, according to Equations 1 to 5.

Reformulating Existing Asymptotic Person Fit Indices

This section summarizes how I_z , U , W , ECI2_z, and ECI4_z can be expressed in both Forms 1 and 4. Their corresponding asymptotic versions in Form 5, I_z^* , U^* , W^* , ECI2_z^{*}, and ECI4_z^{*}, can thus be calculated. To avoid repetition of the formulas, this section does not show the transformation again from the centered Form 4 to the asymptotic Form 5. The only mathematical differences among the above indices lie in $v_n(\hat{\theta})$ (and thus $\omega_n(\hat{\theta})$). Online Appendix A illustrates $v_n(\hat{\theta})$ for each index as a function of $P_n(\hat{\theta})$.

I_z and I_z^*

I_z (Drasgow et al., 1985) is a special case when

$$v_n(\hat{\theta}) = \frac{1}{Q_n(\hat{\theta}) - P_n(\hat{\theta})} \log \frac{P_n(\hat{\theta})}{Q_n(\hat{\theta})} \quad (6)$$

and thus $\omega_n(\hat{\theta}) = \log(P_n(\hat{\theta})/Q_n(\hat{\theta}))$. Snijders (2001) developed its asymptotic version I_z^* when the estimated ability parameter is used. As shown in Online Appendix A, Equation 6 is a symmetric weight function with respect to $P_n(\hat{\theta}) = .5$, and it gives larger weights (in absolute values) to items with $P_n(\hat{\theta})$ approaching 1 and 0, and smaller weights to items with $P_n(\hat{\theta})$ approaching .5. Therefore, I_z^* is expected to be equally attentive to detecting spuriously high scores on difficult items and spuriously low scores on easy items. As Equation 6 is always a negative value, an I_z^* falling below a negative critical value (e.g., $I_z^* < -1.64$ at $\alpha = .05$) indicates an aberrant response pattern.

U and U^*

Wright and Stone (1979; see also Wright & Masters, 1982) proposed U as a residual-based person fit statistic. U is derived by using a weight function in Form 1 as

$$v_n(\hat{\theta}) = \frac{1}{NP_n(\hat{\theta})(1 - P_n(\hat{\theta}))}. \quad (7)$$

As mentioned above, multiplying a constant with $v_n(\hat{\theta})$ does not change U^* . According to Equation 3 and dividing Equation 7 by the constant $1/N$, $\omega_n(\hat{\theta}) = 1/P_n(\hat{\theta}) - 1/Q_n(\hat{\theta})$. The asymptotic version, U^* (Magis et al., 2014), is computed using Equations 5 and 7. The weight function (Equation 7) is symmetric with respect to $P_n(\hat{\theta}) = .5$, as illustrated in Online Appendix A. U^* also gives larger weights to the items with $P_n(\hat{\theta})$ approaching 1 or 0, and the least weights to the items with $P_n(\hat{\theta}) = .5$. U^* is therefore expected to be equally attentive to detecting

spuriously high scores on difficult items and spuriously low scores on easy items. Because Equation 7 is always positive, a value of U^* exceeding the positive critical value from the standard normal distribution (e.g., $U^* > 1.64$ at $\alpha = .05$) indicates an aberrant response pattern.

*W and W^**

Wright and Stone (1979; see also Wright & Masters, 1982) also proposed W as another person fit statistic. W uses the weight

$$v_n(\hat{\theta}) = \frac{1}{\sum_{n=1}^N P_n(\hat{\theta})(1 - P_n(\hat{\theta}))}, \quad (8)$$

which is a constant value across all the items (i.e., it weights all items equally). Therefore, Equation 8 can be simplified as a constant weight 1. Its weight in the centered form becomes $\omega_n(\hat{\theta}) = Q_n(\hat{\theta}) - P_n(\hat{\theta})$. Because of the constant weight in Equation 8 as shown in Online Appendix A, its ability to detect spuriously high scores on the most difficult items (e.g., items with $P_n(\hat{\theta}) < .10$) and spuriously low scores on the easiest items (e.g., items with $P_n(\hat{\theta}) > .90$) are expected to be lower than I_z^* and U^* . As Equation 8 is always positive, a value of W^* exceeding the critical values obtained from the standard normal distribution indicates person misfit (e.g., $W^* > 1.64$ at $\alpha = .05$).

ECI2_z, ECI2_z^{}, ECI4_z, and ECI4_z^{*}*

Sato (1975, 1982), Tatsuoaka and Tatsuoaka (1982), and Tatsuoaka (1984) proposed the so-called caution index, several versions of extended caution indices, and the corresponding standardized extended caution indices (ECI_z). Among them, ECI2_z^{*} and ECI4_z^{*} have been the most popular and have been generalized to ECI2_z^{*} and ECI4_z^{*} by Sinharay (2016a) using Snijders's (2001) correction.

Let M be the total number of observations, and m be the m th observation. $T(\hat{\theta})$, G_n , and G are defined as $T(\hat{\theta}) = \sum_{n=1}^N P_n(\hat{\theta})/N$, $G_n = \sum_{m=1}^M P_n(\hat{\theta}_m)/M$, and $G = \sum_{m=1}^M \sum_{n=1}^N P_n(\hat{\theta}_m)/(M \cdot N)$. Conceptually, $T(\hat{\theta})$ is the average probability of correct answer across the N items for a given $\hat{\theta}$, G_n is the n th item's average probability of correct answer across the M observations, and G is the overall average.

For ECI2_z^{*},

$$v_n(\hat{\theta}) = \frac{-(G_n - G)}{Q_n(\hat{\theta}) - P_n(\hat{\theta})} \quad (9)$$

and $\omega_n(\hat{\theta}) = -(G_n - G)$. For ECI4_z^{*},

$$v_n(\hat{\theta}) = \frac{-(P_n(\hat{\theta}) - T(\hat{\theta}))}{Q_n(\hat{\theta}) - P_n(\hat{\theta})} \quad (10)$$

and $\omega_n(\hat{\theta}) = -(P_n(\hat{\theta}) - T(\hat{\theta}))$. Unlike I_z^* , U^* , and W^* , the weights in Equations 9 and 10 are not only functions of $P_n(\hat{\theta})$, but also impacted by $T(\hat{\theta})$, G_n , and G . More specifically, both $\hat{\theta}$ and item parameters can alter the relative weights of different items. The second and third rows in Online Appendix A illustrate the shapes of the weights in Equations 9 and 10 with $\hat{\theta}$ at -2 , -1 , and 0 . In this example,² when $\hat{\theta}$ is 0 , both ECI2_z^{*} and ECI4_z^{*} have symmetric weights, and

ECI2_z* attach heavier weights to items with more extreme difficulty levels. When $\hat{\theta}$ is -1 or -2 , their weights are asymmetric and attach heavier weights to items whose $P_n(\hat{\theta})$ are closer to .5. The authors therefore speculate that the sensitivity of ECI2_z* and ECI4_z* to person misfit may change dramatically depending on person and item parameters.

Introducing SHa(λ)* and SHb(β)*

Snijders's (2001) framework is flexible in that by specifying the weight functions appropriately, the various person fit indices uniformly follow the standard normal distribution asymptotically when estimated ability parameters are used instead of the true values. Because I_z^* , U^* , and W^* (as well as ECI2_z* and ECI4_z* given some specific person and item parameters) have symmetric weight functions in Form 1, they split their power equally between identifying spuriously high scores on difficult items and identifying spuriously low scores on easy items. The authors hereby propose SHa(λ)* and SHb(β)* with the aim of detecting spuriously high scores on difficult items specifically. To achieve this goal, heavier weights are assigned to items with $P_n(\hat{\theta})$ closer to 0 in SHa(λ)* and SHb(β)*.

Specifically, for SHa(λ)*,

$$v_n(\hat{\theta}) = \frac{1}{P_n^\lambda(\hat{\theta})},$$

and thus, $\omega_n(\hat{\theta}) = (Q_n(\hat{\theta}) - P_n(\hat{\theta})) / P_n^\lambda(\hat{\theta})$. The weight distribution of SHa(λ)* varies by the value of λ (>0). For example, as shown in Online Appendix A, SHa(1/2)* (i.e., $\lambda = 1/2$) starts to have large weights when $P_n(\hat{\theta})$ falls below .2, and large weights in SHa(1)* (i.e., $\lambda = 1$) mainly locate on the more difficult items with $P_n(\hat{\theta}) < .1$. The authors therefore expect that SHa(1/2)* and SHa(1)* can be very sensitive to spuriously high scores on items with $P_n(\hat{\theta}) < .2$ and $P_n(\hat{\theta}) < .1$, respectively, but may lose power for aberrant responses on those with $P_n(\hat{\theta})$ between .2 and .3.

To have acceptable power to detect spuriously high scores on a wider range of difficult items (our targets are the items with $P_n(\hat{\theta}) < .3$), the authors suggest SHb(β)* with the following weight:

$$v_n(\hat{\theta}) = \exp\left(-(\beta \cdot P_n(\hat{\theta}))^2\right),$$

and thus, $\omega_n(\hat{\theta}) = (Q_n(\hat{\theta}) - P_n(\hat{\theta})) \cdot \exp\left(-(\beta \cdot P_n(\hat{\theta}))^2\right)$. As shown in Online Appendix A, SHb(2)* (i.e., $\beta = 2$) and SHb(3)* (i.e., $\beta = 3$) have relatively large weights on the items with $P_n(\hat{\theta}) < .3$ compared with SHa(1/2)* and SHa(1)*. The authors expect that SHb(2)* and SHb(3)* can have greater sensitivity to detect spuriously high scores on items with $.1 < P_n(\hat{\theta}) < .3$, although they may lose some power for the extremely difficult items with $P_n(\hat{\theta}) < .1$ compared with SHa(1/2)* and SHa(1)*.

SHa(λ)* and SHb(β)* exceeding the critical value from the standard normal distribution (e.g., >1.64 at $\alpha = .05$) indicate spuriously high scores on difficult items. By using an asymmetric weight functions, they are expected to result in improved power of detecting spuriously high scores on difficult items as compared with I_z^* , U^* , and W^* . Because the weights of ECI2_z* and ECI4_z* depend on the person and item parameters, their performance relative to the other person fit indices may differ across conditions. Online Appendix B provides a simple-case demonstration of the mechanism of SHa(λ)* and SHb(β)* in detecting spuriously high scores on difficult items. The study's simulation results explored different values of λ and β . The authors found that $\lambda = 1/2$ and 1 and $\beta = 2$ and 3 provide the most meaningful results in terms of power. In contrast, $\lambda = 2$ and $\beta = 4$, for instance, resulted in very low power ($<5\%$) when

detecting items with $.05 < P_n(\hat{\theta}) < .3$, although they led to strong power of detecting items with $P_n(\hat{\theta}) < .05$. Therefore, the following simulation studies will only present results from SHa(1/2)*, SHa(1)*, SHb(2)*, and SHb(3)*. In addition, $\omega_n(\theta)$ for each proposed index and its first and second derivatives are all uniformly bounded in a neighborhood of θ , satisfying Snijders's constraints on $\omega_n(\theta)$.

Simulation Study I: Type I Error Rates

This section presents a simulation study that investigated the empirical Type I error rates under the null distribution, where responses were simulated using the 2PLM and no aberrant response existed. The authors varied θ at $-2, -1, 0, 1$, and 2 . Ideally, a good person fit index should yield well-controlled Type I error rates across various θ values. The test length n varied at 10 (extremely short test), 20 (short test), and 50 (moderate length). Because the standard normal distribution of the above person fit indices is based on asymptotic theory, which requires as many test items as possible, the authors expect and will show that $n = 10$ can lead to inaccurate Type I error rates in some conditions.

For the item parameters, when n was 10, a took the value of $[.7, 1.3]$ and crossed with b parameters at $[-1.8, -.9, 0, .9, 1.8]$. When n was 20, a took the value of $[.7, .9, 1.1, 1.3]$ and crossed with b at $[-1.8, -.9, 0, .9, 1.8]$. When n was 50, a took the value of $[.7, .9, 1.0, 1.1, 1.3]$ and crossed with b ranging from -1.8 to 1.8 with an interval of $.4$.

For each of the 15 conditions (5θ values $\times 3$ test length), the authors simulated 100,000 replications. MLE was employed to estimate the ability parameter. Type I error rates were evaluated at two nominal levels: $\alpha = .05$ or $.01$. The simulation study was implemented in R 3.3.2 (R Development Core Team, 2017), and the ability parameter was estimated using the TAM package (Kiefer, Robitzsch, & Wu, 2017).

Table 1 shows the results for the Type I error rates. Because higher Type I error rates pose more serious concerns than lower rates, the values 1% greater than the target value in Table 1 were made bold. Across all the conditions, only $ECI4_z^*$ had the most well-controlled Type I error rates. When $n = 10$, I_z^* , $ECI2_z^*$, SHa(1/2)*, SHa(1)*, SHb(2)*, and SHb(3)* all tended to yield inflated Type I error rates, especially for SHb(3)* (8.1% and 14.0%) with $\alpha = 5\%$ and $\theta = 1$ and 2 . When $n = 20$, U^* and SHa(1)* had the most numbers of conditions with inflated Type I error rates, but SHb(2)* and SHb(3)* had the largest values (7.2% and 7.4%, respectively) at $\alpha = 5\%$ and $\theta = 2$. However, when $n = 50$, the majority of the person fit indices tended to have well-controlled Type I error rates except for most of the values from U^* , and a few from SHa(1/2)* (when $\theta = -2$) and SHa(1)* (when $\theta = -2$ and -1).

Simulation Study II: Power for Detecting Spuriously High Scores

The authors implemented a second simulation study to investigate the power of detecting spuriously high scores on difficult items. The simulation design was similar to the design of Simulation Study I, except that aberrant responses were simulated for some items. In addition, the authors limited θ at $-2, -1$, and 0 to mimic the examinees who have relatively low ability and thus can potentially have spuriously high scores on difficult items.

The authors manipulated the percentage of the items that had aberrant responses at 10% or 20%. Two scenarios of aberrant responses were generated. In Scenario 1, the 10% (or 20%) most difficult items (i.e., the lowest $P_n(\theta)$) were always answered correctly, which mimicked the case where a person has preknowledge of the most difficult items. Scenario 2 represented the situation where the person knows the moderately difficult items but does not know the most difficult ones. Thus, in Scenario 2, the items with $P_n(\theta)$ between the 10th and 20th percentile (or 20th to 40th

Table 1. Type I Error Rates (%) Under the Null Distribution With No Aberrant Response.

α	1%					5%				
θ	-2	-1	0	1	2	-2	-1	0	1	2
$n = 10$										
I_z^*	2.7	2.1	1.9	2.2	2.8	6.9	5.4	5.0	5.6	7.1
U^*	1.3	1.9	2.4	2.1	1.3	1.9	3.3	4.5	3.4	2.0
W^*	1.6	1.8	1.8	1.8	1.7	7.0	5.8	5.5	6.0	7.3
$ECI2_z^*$	2.6	2.0	1.7	2.0	2.7	6.8	5.5	5.3	5.7	6.9
$ECI4_z^*$	0.9	1.5	1.6	1.5	1.0	4.5	5.0	5.3	5.3	4.5
$SHa(1/2)^*$	1.9	2.3	2.3	2.1	1.8	4.3	4.8	5.6	6.2	7.3
$SHa(1)^*$	1.6	2.4	3.2	2.5	1.7	1.8	3.1	5.7	7.0	7.1
$SHb(2)^*$	2.6	2.1	2.1	1.7	0.8	6.7	5.6	5.6	6.9	8.1
$SHb(3)^*$	2.6	2.2	2.3	1.7	0.6	6.5	6.2	7.9	9.4	14.0
$n = 20$										
I_z^*	2.2	1.7	1.6	1.7	2.2	6.2	5.5	5.1	5.3	6.2
U^*	2.4	2.8	2.4	2.7	2.5	4.0	5.3	5.7	5.2	4.0
W^*	1.6	1.5	1.5	1.4	1.7	6.2	5.3	4.9	5.1	6.4
$ECI2_z^*$	2.3	1.7	1.5	1.6	2.3	6.3	5.5	5.0	5.3	6.3
$ECI4_z^*$	1.2	1.4	1.5	1.4	1.2	5.1	5.2	4.9	5.1	5.2
$SHa(1/2)^*$	2.7	2.3	1.7	1.4	1.6	5.8	5.8	5.3	5.2	6.1
$SHa(1)^*$	2.5	3.1	2.5	1.5	1.6	4.0	5.6	6.1	5.5	6.0
$SHb(2)^*$	1.8	1.7	1.6	1.3	0.8	6.3	5.6	5.4	5.4	7.2
$SHb(3)^*$	1.9	1.8	1.9	1.4	0.6	6.3	5.9	5.7	5.8	7.4
$n = 50$										
I_z^*	1.6	1.4	1.3	1.5	1.6	5.7	5.3	5.1	5.2	5.8
U^*	2.9	2.6	1.7	2.5	2.9	5.6	6.0	5.5	6.0	5.7
W^*	1.3	1.3	1.3	1.4	1.4	5.5	5.1	5.0	5.1	5.4
$ECI2_z^*$	1.7	1.4	1.3	1.5	1.7	5.7	5.2	5.0	5.3	5.9
$ECI4_z^*$	1.1	1.2	1.3	1.3	1.2	5.0	5.0	5.0	5.0	5.1
$SHa(1/2)^*$	2.3	1.7	1.4	1.4	1.3	6.1	5.5	5.3	5.1	5.5
$SHa(1)^*$	3.0	2.9	1.9	1.4	1.3	5.6	6.3	5.8	5.2	5.4
$SHb(2)^*$	1.3	1.3	1.4	1.3	0.9	5.5	5.2	5.2	5.4	5.3
$SHb(3)^*$	1.4	1.4	1.6	1.4	1.1	5.5	5.4	5.6	5.7	5.9

Note. When $\alpha = 1\%$, the values $\geq 2.0\%$ were made bold. When $\alpha = 5\%$, the values $\geq 6.0\%$ were made bold. $\alpha =$ nominal alpha level; $\theta =$ true value of the ability parameter; $n =$ test length.

percentile) were always answered correctly. The authors used “1st—10%,” “2nd—10%,” “1st—20%,” and “2nd—20%” to denote the above conditions. Online Appendix C presents the $P_n(\theta)$ of the items that had spuriously high scores. For each n and θ , $P_n(\theta)$ was ranked from the lowest (the most difficult) to the highest (the least difficult). For example, when $n = 50$, the items with rank 1 to 5 had spuriously high scores under the 1st—10% condition, and the items with rank 6 to 10 had spuriously high scores under the 2nd—10% condition.

Table 2 presents the empirical power of each person fit index under the 1st—10% condition, where $P_n(\theta)$ ranged from .070 to .121 (very difficult items). For each condition, the highest three detection rates (and at least 20.0%) or the values greater than 90.0% were made bold in Table 2. $SHa(1/2)^*$ and $SHa(1)^*$ had the largest power in most of the conditions, compared with the other indices. $SHb(2)^*$ and $SHb(3)^*$ yielded relatively large power when the test length reached 20. U^* also had large power, especially when $\theta = -1$ and -2 , but not as much as $SHa(1/2)^*$ and $SHa(1)^*$. The detection rates of I_z^* , $ECI2_z^*$, $ECI4_z^*$, and especially W^* were generally 10% to 40% lower than $SHa(1/2)^*$ and $SHa(1)^*$.

Table 2. Power (%) for Detecting Spuriously High Scores Under the 1st—10% Condition.

α	1%			5%		
θ	0	-1	-2	0	-1	-2
<i>n</i> = 10						
I_z^*	13.1	27.0	51.2	29.2	50.5	76.3
U^*	18.1	57.9	90.6	29.5	72.6	96.1
W^*	11.0	14.3	15.2	28.0	35.4	41.3
$ECI2_z^*$	10.4	19.3	33.1	26.1	45.0	70.7
$ECI4_z^*$	9.9	15.2	20.6	26.0	35.6	42.4
$SHa(1/2)^*$	22.6	55.9	87.1	47.0	81.0	97.3
$SHa(1)^*$	35.6	77.3	97.0	61.4	92.2	99.5
$SHb(2)^*$	17.1	24.8	25.5	39.9	49.4	48.2
$SHb(3)^*$	19.9	28.5	26.1	70.9	92.0	81.5
<i>n</i> = 20						
I_z^*	23.3	49.5	82.4	46.9	75.8	95.5
U^*	25.6	77.5	99.0	40.1	88.1	99.7
W^*	20.7	28.7	33.8	45.6	56.1	62.8
$ECI2_z^*$	18.4	40.0	74.3	39.4	67.2	92.4
$ECI4_z^*$	19.4	31.1	45.1	43.0	59.3	75.8
$SHa(1/2)^*$	38.5	81.0	98.7	67.6	95.3	99.9
$SHa(1)^*$	55.9	95.4	100.0	81.0	99.2	100.0
$SHb(2)^*$	40.5	53.8	50.7	81.1	94.7	92.6
$SHb(3)^*$	60.0	85.3	68.7	90.2	99.8	100.0
<i>n</i> = 50						
I_z^*	54.7	89.2	99.6	80.5	98.0	100.0
U^*	45.3	97.5	100.0	66.7	99.4	100.0
W^*	54.8	70.9	80.5	82.1	91.7	96.4
$ECI2_z^*$	45.5	82.2	98.9	72.4	95.7	99.9
$ECI4_z^*$	52.3	74.0	91.5	79.8	93.1	98.9
$SHa(1/2)^*$	82.4	99.5	100.0	96.5	100.0	100.0
$SHa(1)^*$	94.8	100.0	100.0	99.4	100.0	100.0
$SHb(2)^*$	98.7	99.6	99.2	99.9	100.0	100.0
$SHb(3)^*$	99.7	100.0	100.0	99.9	100.0	100.0

Note. The highest three values that are at least 20.0% and all the values greater than 90% under each condition are made bold. α = nominal alpha level; θ = true value of the ability parameter; *n* = test length.

Table 3 presents the empirical power under the 2nd—10% condition, with $P_n(\theta)$ ranging from .023 to .221 (very difficult to moderately difficult). With *n* = 10, the detection rates were generally low, except for $SHa(1/2)^*$ (30.9%) when $\theta = -2$ and $\alpha = 1\%$, and for $SHa(1/2)^*$ (47.1%), I_z^* (45.8%), and U^* (38.2%) when $\theta = -2$ and $\alpha = 5\%$. With *n* = 20 and 50, $SHa(1/2)^*$, $SHb(2)^*$, and $SHb(3)^*$ led to the highest detection rates in most conditions. The detection rates of $SHa(1)^*$ were large only when $\theta = -2$. I_z^* , U^* , W^* , $ECI2_z^*$, and $ECI4_z^*$ performed well in some conditions (e.g., when $\theta = -2$ and *n* = 50) but had lower power in general than the authors' proposed indices.

Table 4 summarizes the empirical power under the 1st—20% condition, with $P_n(\theta)$ ranging from .007 to .221 (from very difficult to moderately difficult). Across all the conditions, $SHa(1/2)^*$, $SHa(1)^*$, $SHb(2)^*$, and $SHb(3)^*$ outperformed the other indices, especially when *n* = 10 and 20 and $\theta = -1$. W^* and $ECI4_z^*$ also appeared to have large power when *n* = 20 across all θ values.

Table 5 presents the empirical power under the 2nd—20% condition. In this condition, the values of $P_n(\theta)$ ranged from .04 to .348 and thus represented more moderate levels of difficulty compared with the other conditions. When *n* = 10, the detection rates were generally small, except for

Table 3. Power (%) for Detecting Spuriously High Scores Under the 2nd—10% Condition.

α	1%			5%		
θ	0	−1	−2	0	−1	−2
$n = 10$						
I_z^*	4.9	8.8	16.1	11.1	23.0	45.8
U^*	3.9	5.9	14.1	6.9	16.9	38.2
W^*	5.0	7.8	10.9	13.6	20.8	28.6
$ECI2_z^*$	5.0	8.3	15.1	13.7	21.2	33.7
$ECI4_z^*$	4.5	9.0	16.1	13.0	22.3	33.7
$SHa(1/2)^*$	4.7	10.3	30.9	10.8	21.3	47.1
$SHa(1)^*$	4.6	5.9	22.9	7.1	11.2	34.7
$SHb(2)^*$	6.6	9.1	15.6	13.4	19.6	27.2
$SHb(3)^*$	6.7	9.3	16.6	11.9	18.9	33.1
$n = 20$						
I_z^*	8.9	20.8	49.4	22.7	43.9	76.7
U^*	7.6	17.0	56.4	17.0	41.7	83.9
W^*	8.9	17.0	27.5	23.6	39.0	55.2
$ECI2_z^*$	9.8	20.8	48.3	25.2	44.3	77.3
$ECI4_z^*$	9.3	17.9	30.9	24.6	40.3	60.4
$SHa(1/2)^*$	9.6	23.2	63.4	24.5	49.2	87.2
$SHa(1)^*$	8.2	13.8	54.6	18.2	37.2	84.8
$SHb(2)^*$	12.8	23.8	38.2	30.9	50.6	71.1
$SHb(3)^*$	12.4	26.6	45.2	27.3	57.8	93.0
$n = 50$						
I_z^*	21.0	55.3	91.5	44.5	81.3	98.7
U^*	16.2	54.9	97.2	34.2	81.7	99.7
W^*	22.6	48.5	71.4	47.7	76.6	92.2
$ECI2_z^*$	22.0	56.1	91.7	46.3	81.9	98.8
$ECI4_z^*$	22.9	49.3	78.2	48.3	77.3	95.0
$SHa(1/2)^*$	22.7	66.0	97.8	48.5	89.5	99.9
$SHa(1)^*$	16.7	56.1	98.2	39.1	87.3	99.9
$SHb(2)^*$	26.1	73.9	92.4	52.5	95.0	99.8
$SHb(3)^*$	20.4	81.6	99.5	42.3	96.8	100.0

Note. The highest three values that are at least 20.0% and all the values greater than 90% under each condition are made bold. α = nominal alpha level; θ = true value of the ability parameter; n = test length.

$SHb(2)^*$, $SHb(3)^*$, W^* , $ECI2_z^*$, and $ECI4_z^*$ with $\theta = -2$. When $n = 20$, the detection rates of W^* , $ECI2_z^*$, $ECI4_z^*$ were similarly large. $SHb(2)^*$ also resulted in acceptable power, about only 1% lower than W^* , $ECI2_z^*$, $ECI4_z^*$ when $\theta = -1$ and -2 . However, $SHb(2)^*$ lost the power when $\theta = 0$. When n reached 50, W^* , $ECI2_z^*$, and $ECI4_z^*$ also had similar values that were larger than the other indices. I_z^* , followed by $SHb(2)^*$, had the next highest power when $\theta = -1$ and -2 .

Conclusions and Discussion

This article proposes two new person fit indices, $SHa(\lambda)^*$ and $SHb(\beta)^*$, that are designed to focus their power on detecting spuriously high scores. The development of these two indices was motivated by a common scenario that has not been directly addressed by the existing person fit indices: In many educational tests, the purpose of the test is to select individuals with high ability levels. Thus, spuriously high scores are a more serious threat to the validity of the test, and also there tend to be more attempts to achieve spuriously high scores than spuriously low

Table 4. Power (%) for Detecting Spuriously High Scores Under the 1st—20% Condition.

α	1%			5%		
θ	0	-1	-2	0	-1	-2
$n = 10$						
I_z^*	28.6	50.7	78.1	51.3	79.0	96.0
U^*	18.8	45.7	81.6	31.3	62.5	91.4
W^*	28.9	51.5	77.9	59.4	79.0	96.0
$ECI2_z^*$	25.7	45.4	74.5	51.2	69.9	91.2
$ECI4_z^*$	28.1	45.8	74.9	59.1	75.3	95.4
$SHa(1/2)^*$	44.1	65.3	92.4	71.2	85.5	98.3
$SHa(1)^*$	52.4	69.7	94.7	76.6	84.7	98.2
$SHb(2)^*$	53.6	61.8	91.9	80.7	76.4	96.6
$SHb(3)^*$	57.3	57.4	90.3	81.8	83.6	98.2
$n = 20$						
I_z^*	60.8	89.8	99.5	84.4	97.9	100.0
U^*	32.8	76.2	98.5	49.6	88.0	99.6
W^*	70.6	92.4	99.6	92.7	99.0	100.0
$ECI2_z^*$	56.7	85.3	99.0	80.7	95.7	99.9
$ECI4_z^*$	72.5	89.6	99.4	94.4	98.1	100.0
$SHa(1/2)^*$	82.8	97.1	100.0	95.6	99.6	100.0
$SHa(1)^*$	88.3	98.2	100.0	97.3	99.8	100.0
$SHb(2)^*$	89.3	97.9	100.0	93.9	99.6	100.0
$SHb(3)^*$	88.6	98.4	100.0	91.0	99.7	100.0
$n = 50$						
I_z^*	96.4	100.0	100.0	99.6	100.0	100.0
U^*	66.0	98.9	100.0	83.4	99.8	100.0
W^*	99.3	100.0	100.0	100.0	100.0	100.0
$ECI2_z^*$	94.2	99.9	100.0	99.1	100.0	100.0
$ECI4_z^*$	98.7	100.0	100.0	99.9	100.0	100.0
$SHa(1/2)^*$	99.7	100.0	100.0	100.0	100.0	100.0
$SHa(1)^*$	99.8	100.0	100.0	100.0	100.0	100.0
$SHb(2)^*$	98.2	100.0	100.0	99.1	100.0	100.0
$SHb(3)^*$	95.8	100.0	100.0	97.7	100.0	100.0

Note. The highest three values that are at least 20.0% and all the values greater than 90% under each condition are made bold. α = nominal alpha level; θ = true value of the ability parameter; n = test length.

scores. I_z^* , U^* , and W^* cannot provide optimal detection power in the above scenario because they rely on symmetric weight functions, which distributes the detection power equally between detecting spuriously high scores and spuriously low scores. The weight functions of $ECI2_z^*$ and $ECI4_z^*$ have different shapes, contingent upon person and item parameters, which can impact the detection power in a way that are currently not fully aware. $SHa(\lambda)^*$ and $SHb(\beta)^*$, on the contrary, utilize asymmetric weight functions and place heavier weights on items that are more difficult for an examinee. Consequently, when test administrators want to monitor spuriously high scores primarily, the proposed indices will provide more powerful targeted detection. Two candidate weight functions were proposed, along with two candidate values (i.e., $\lambda = 1/2$ or 1 and $\beta = 2$ or 3) for their tuning parameters.

Simulation Study I finds that the Type I error rates are generally acceptable when test length is 20 or 50. $ECI4^*$ has the best Type I error rates across all conditions. U^* has inflated Type I error rates across most θ s even when n reaches 50. However, when $n = 10$, all the indices except for $ECI4^*$ have several inflated Type I error rates, especially for $SHb(3)^*$ when $\alpha = 5\%$ and $\theta = -2$. Consistent with previous findings (Molenaar & Hoijsink, 1990; Nering, 1995; Reise,

Table 5. Power (%) for Detecting Spuriously High Scores Under the 2nd—20% Condition.

α	1%			5%		
θ	0	-1	-2	0	-1	-2
<i>n</i> = 10						
I_z^*	6.2	15.9	20.9	14.7	33.2	49.1
U^*	6.0	4.8	2.4	10.1	10.2	7.7
W^*	6.0	16.9	27.6	16.0	33.4	52.5
$ECI2_z^*$	6.3	18.3	25.3	17.2	40.4	57.1
$ECI4_z^*$	6.1	11.5	23.9	16.5	26.1	57.0
$SHa(1/2)^*$	4.0	8.1	9.6	10.2	25.4	40.2
$SHa(1)^*$	1.9	2.7	1.0	4.6	4.0	1.6
$SHb(2)^*$	3.0	17.0	31.3	9.4	45.4	53.7
$SHb(3)^*$	2.6	18.9	27.6	7.9	37.3	53.3
<i>n</i> = 20						
I_z^*	9.6	31.8	65.8	23.1	58.7	86.9
U^*	10.5	11.6	14.0	19.5	34.7	60.9
W^*	9.5	36.4	70.2	23.4	64.4	89.7
$ECI2_z^*$	10.9	38.3	69.5	26.1	67.0	90.2
$ECI4_z^*$	9.7	33.3	70.6	24.0	62.9	89.9
$SHa(1/2)^*$	4.9	23.2	53.2	14.1	50.6	81.6
$SHa(1)^*$	2.1	5.9	5.7	6.7	16.3	32.9
$SHb(2)^*$	2.9	37.5	71.4	7.7	61.8	88.3
$SHb(3)^*$	1.5	29.0	65.2	4.2	48.8	82.8
<i>n</i> = 50						
I_z^*	17.9	65.3	97.2	38.1	87.1	99.7
U^*	16.8	39.1	78.5	33.2	70.0	97.3
W^*	18.1	72.9	98.4	38.5	90.8	99.8
$ECI2_z^*$	22.2	72.8	98.4	44.7	91.2	99.8
$ECI4_z^*$	20.3	72.9	98.3	42.1	91.2	99.8
$SHa(1/2)^*$	9.1	50.3	94.9	24.1	78.8	99.5
$SHa(1)^*$	3.4	13.6	48.5	11.1	37.7	89.1
$SHb(2)^*$	2.6	52.0	97.8	8.0	73.9	99.7
$SHb(3)^*$	0.7	30.8	94.0	2.7	51.1	98.6

Note. The highest three values that are at least 20.0% and all the values greater than 90% under each condition are made bold. α = nominal alpha level; θ = true value of the ability parameter; *n* = test length.

1995; Snijders, 2001), a further examination of the indices reveals that their null distributions are highly skewed when *n* = 10. Therefore, those indices need to be used with caution for very short tests. Resampling methods (e.g., bootstrap) or critical values from simulations that yield a target false-positive rate may potentially help control Type I error for extremely short tests but require further study.

As for the proposed indices, although they had inflated Type I error rates in a few conditions even when the test was longer (i.e., 20 items or 50 items), this is deemed less of concern for two reasons. First, the resulting empirical Type I error rates are only slightly greater than the specified value (i.e., within 3% Type I error rates inflation). The improved sensitivity to detect spurious high scores (at least 20% increase in power for the authors' proposed indices under most conditions) overshadows the scale of the potential inflation in. Second, in the setting of aberrant response detection, Type I error is less harmful than Type II error. In other words, person fit indices are used as a screening tool to identify items that are potentially breached out of a large pool of items. Then, human review and further investigation should be done to determine the action. When a Type I error is made, a clean item is flagged, but further investigation

can recover the item and deem it is still clean. Compared with that, it is more dangerous that the screening tool misses a breached item (i.e., having less power) and that item stays in the test and continues to damage the validity of the test for all examinees. Thus, an index with greater power is strongly preferred in this setting.

In this simulation study, the authors empirically confirmed that the proposed indices generally had greater power for the specific scenario being discussed: spuriously high scores for difficult items. The authors investigated $\lambda = .5, 1$ and $\beta = 2$ and 3 for $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$, respectively. In general, larger values of λ and β result in heavier weights and thus greater detection power for aberrant responses on more difficult items, at the cost of less power for less difficult items. Practitioners could adjust the values of λ and β to target their own research problems. Specifically, the performance of the person fit indices also depends on the difficulty of the breached items given a person's ability. When the spuriously high scores occur on very difficult items (e.g., the items with $P_n(\theta) < 10.0\%$, corresponding to the 1st—10% and 1st—20% conditions with $\theta = -1$ and -2), $\text{SHa}(1/2)^*$, $\text{SHa}(1)^*$, $\text{SHb}(2)^*$, and $\text{SHb}(3)^*$ generally have the highest power, followed by U^* . The results are not surprising because $\text{SHa}(1/2)^*$, $\text{SHa}(1)^*$, $\text{SHb}(2)^*$, $\text{SHb}(3)^*$, and U^* all have relatively large weights when $P_n(\theta) < 10.0\%$ as shown in Online Appendix A. On the contrary, when the moderate items (e.g., the items with $10.0\% < P_n(\theta) < 30.0\%$, corresponding to the 2nd—20% condition with $\theta = 0$ and 1) have spuriously high scores, W^* , ECI2_z^* , and ECI4_z^* have the best detection rates.

As a summary, the authors provide the following recommendations when applying the person fit indices considered in this study to detect spuriously high scores:

1. All the indices need to be applied with caution under extremely short tests (e.g., $n = 10$) because of the pervasive Type I error inflation.
2. When the purpose is to detect spuriously high scores on very difficult items (e.g., items with smaller than 10% chances of correct responses for examinees with low ability levels), $\text{SHa}(1/2)^*$ and $\text{SHb}(3)^*$ have superior power to the other indices. However, $\text{SHb}(3)^*$ can only be applied when n is at least 20 due to its highly inflated Type I error rates when $n = 10$ and $\theta = -1$ or -2 .
3. When the purpose is to detect spuriously high scores on items with moderate to high levels of difficulty (e.g., $.10 < P_n(\theta) < .30$), W^* can be a good choice. $\text{SHb}(2)^*$ may also be considered because it maintains some power in this situation and also has acceptable power when detecting more difficult items.
4. When both spuriously high and low scores are of interest, I_z^* , U^* , and W^* should be considered. Although the above simulations do not concern such a situation, I_z^* , U^* , and W^* have symmetric weights, and thus, each of them has equal sensitivity to a correct answer on a difficult item (e.g., with $P_n(\theta) = .10$) and a wrong answer on an easy item (e.g., $P_n(\theta) = .90$).

For ECI2_z^* and ECI4_z^* , this study has not thoroughly investigated the shapes of their weight functions. Because the shapes of ECI2_z^* and ECI4_z^* can differ, depending on θ and item parameters, they may be sensitive to detect spuriously high or low scores in some circumstances, but not in others. Online Appendix D illustrates the weight functions of ECI2_z^* and ECI4_z^* based on the person and item parameters used in Simulation Study II. Both of ECI2_z^* and ECI4_z^* have symmetric weights when $\theta = 0$. When $\theta = -1$ and -2 , items with $P_n(\theta) < .3$ have larger weights than those with $P_n(\theta) < .5$. However, the items with $P_n(\theta) > .5$ have surprisingly the largest weights when $\theta = -1$, and the authors are not aware of the rationale behind this phenomenon. Therefore, the application of ECI2_z^* and ECI4_z^* requires further investigations.

In practice, the population value of $P_n(\theta)$ is unknown. To decide which indices should be used, practitioners can only obtain information on item parameters from either previous calibration or from test developers' expertise in guessing the approximate range of $P_n(\theta)$. When no information exists at all, item parameters need to be estimated first. Aberrant responses are thus expected to bias the item parameter estimates and further impact the detection accuracy regarding person misfit. How to obtain accurate parameter estimates when aberrant responses are present requires further research. For example, Candell and Drasgow's (1988) purification method may be adapted to this setting. A thorough evaluation of the above is needed.

Calculating $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ requires little computational time. Magis et al. (2012) provided the R for I_z^* , but $\text{SHa}(1/2)^*$, $\text{SHa}(1)^*$, $\text{SHb}(2)^*$, and $\text{SHb}(3)^*$ can be similarly computed by modifying the weight functions. Online Appendix E shows how to modify the weight functions using the R codes provided by Magis et al. (2012). As mentioned in the introduction, the proposed indices can be adapted to focus on detecting spuriously low scores instead. The authors use $\text{SLa}(\lambda)^*$ and $\text{SLb}(\beta)^*$ (SL represents spuriously low) to denote the corresponding indices. For $\text{SLa}(\lambda)^*$, $v_n(\hat{\theta}) = 1/Q_n^\lambda(\hat{\theta})$, and thus, $\omega_n(\hat{\theta}) = (Q_n(\hat{\theta}) - P_n(\hat{\theta}))/Q_n^\lambda(\hat{\theta})$. Similarly, for $\text{SLb}(\beta)^*$, $v_n(\hat{\theta}) = \exp(-(\beta \cdot Q_n(\hat{\theta}))^2)$ and $\omega_n(\hat{\theta}) = (Q_n(\hat{\theta}) - P_n(\hat{\theta})) \cdot \exp(-(\beta \cdot Q_n(\hat{\theta}))^2)$.

Simulation studies in this article are limited to the manipulated conditions. Further investigations are needed for generalization purposes. For example, $\hat{\theta}$ is obtained using MLE in this simulation study, but MAP, WLE, expected a posteriori (EAP), the biweight estimate (Mislevy & Bock, 1982), and the Huber estimate (Schuster & Yuan, 2011) can also be employed (see also Sinharay, 2016c). Second, it is of interest to investigate the utility of $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ under 1PLM and 3PLM, both of which are popular models in practice. It is also possible to generalize $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ to models for polytomous items (Snijders, 2001; see also Sinharay, 2016a). In addition, Karabatsos (2003) compared 36 person fit indices and found that four nonparametric person fit indices, that is, H^T (Sijtsma, 1986), C (Sato, (1975), MCI (Harnisch & Linn, 1981), and $U3$ (van der Flier, 1980), outperformed the other parametric statistics in general. Further studies should consider the comparison of $\text{SHa}(\lambda)^*$ and $\text{SHb}(\beta)^*$ with these nonparametric person fit statistics.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Here, the asymptotically standard normal distribution means that as the number of test items increases to infinity, the distribution of I_z^* will converge to the standard normality.
2. For illustration purposes, $T(\hat{\theta})$, G_n , and G were calculated based on a Rasch model with person parameters and difficulty parameters both from -4 to 4 with an interval of .01.

Supplemental Material

Supplementary material is available for this article online.

References

- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12, 253-260.

- 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.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133-146.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298.
- Kiefer, T., Robitzsch, A., & Wu, M. (2017). TAM: Test analysis modules (R package version 1.99993-0) [Computer software manual]. Retrieved from <http://CRAN.R-project.org/web/TAM/index.html>
- Magis, D., Béland, S., & Raïche, 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.
- Magis, D., Raïche, G., & Béland, S. (2012). A didactic presentation of Snijders's I_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, 37, 57-81.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, 42, 725-737.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.
- Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, 9, 27-45.
- Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19, 121-129.
- R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19, 213-229.
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meiji Tokyo.
- Sato, T. (1982). *Application of S-P curve theory*. Tokyo, Japan: Meiji Toshō.
- Schuster, C., & Yuan, K. H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, 36, 720-735.
- Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden*, 7(22), 131-145.
- Sinharay, S. (2016a). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement*, 40, 418-433.
- Sinharay, S. (2016b). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, 81, 992-1013.
- Sinharay, S. (2016c). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of Mathematical and Statistical Psychology*, 69, 175-193.
- Snijders, T. A. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1982). *Standardized extended caution indices and comparison of their error detection rates* (Research Report 82-4-ONR). Urbana: Computer-Based Education Research Laboratory, University of Illinois.
- Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2015). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, 74(5), 1-27.
- van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse, The Netherlands: Swets & Zeitlinger.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago, IL: Mesa Press.