Asymptotic Corrections of Standardized Extended Caution Indices

Applied Psychological Measurement
2016, Vol. 40(6) 418–433
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616649963
apm.sagepub.com

Sandip Sinharay¹

Abstract

Sato and Tatsuoka suggested the caution index, several extended caution indices (ECIs), and their standardized versions. Among these indices, the standardized versions of the second and fourth ECIs (denoted as ζ_1 and ζ_2 , respectively, by Tatsuoka) are arguably the most popular and have been used by several researchers to assess person fit. While Tatsuoka stated that there is satisfactory evidence that ζ_1 and ζ_2 may approximately follow the normal distribution under no person misfit, she also stated that it would be ideal if their theoretical distributions can be derived algebraically. This article derives the “asymptotic” (or “theoretical large-sample”) null distributions of ζ_1 and ζ_2 and suggests corrected versions of them that have asymptotic standard normal null distributions. The derivations are based on the asymptotic correction of person-fit statistics suggested by Snijders. A simulation study shows that the Type I error rate and power of the corrected versions are mostly superior compared with those of the corresponding uncorrected versions. A real data illustration follows. The suggested corrected versions appear to be satisfactory tools for assessing person fit.

Keywords

3PL model, IRT model fit, ζ_1 statistic, ζ_2 statistic

Person-fit assessment (PFA) may help researchers to obtain additional information regarding the answering behavior of persons (Glas & Meijer, 2003). For example, PFA may lead to the conclusion that one’s performance on an intelligence test reflects an unusual lack of concentration on the easiest items instead of one’s true intelligence level. Likewise, in a personality inventory, the PFA of one’s performance may indicate an unusual fear of being evaluated, which is greater or stronger than one’s true level of introversion. In a report for the Council of Chief State School Officers, Olson and Fremer (2013) recommended the use of person-fit statistics (PFSs), in addition to other methods, to detect irregularities in answering behavior. Recent publications such as Ferrando (2015) and Tendeiro and Meijer (2014) described why research on PFA is important.

In the context of item response theory (IRT), several PFSs have been suggested for tests with dichotomous items (e.g., Drasgow, Levine, & Williams, 1985; Klauer, 1991; Smith, 1986; Tatsuoka, 1984; Wright & Stone, 1979). Researchers such as Ferrando (2015), Karabatsos (2003), and Meijer and Sijtsma (2001) provided extensive reviews of PFSs.

¹Pacific Metrics Corporation, Monterey, CA, USA

Corresponding Author:

Sandip Sinharay, Pacific Metrics Corporation, 1 Lower Ragsdale Drive #150, Monterey, CA 93940, USA.
Email: ssinharay@pacificmetrics.com

Sato (1975, 1982) suggested the caution index for detecting unusual response patterns. Tatsuoka and Tatsuoka (1982) and Tatsuoka (1984) suggested several extended caution indices (ECIs) and their standardized versions that have been used as PFSs by several researchers such as Drasgow, Levine, and McLaughlin (1987). Among these statistics, the standardized versions of the second ECI (denoted as ζ_1) and the fourth ECI (ζ_2) are arguably the most popular. However, the null distribution (that refers to the distribution under the null hypothesis of no person misfit) of ζ_1 and ζ_2 is unknown. Tatsuoka stated that it would be ideal if the theoretical distribution of these statistics can be derived algebraically. This article derives the “asymptotic” (or “theoretical large-sample”) null distributions of ζ_1 and ζ_2 and suggests corrected/modified versions of them that follow an asymptotic standard normal null distribution. The derivations are based on the asymptotic correction of PFSs suggested by Snijders (2001).

The “Method” section includes reviews of the ζ_1 and ζ_2 statistics and of the asymptotic correction of Snijders (2001); the section then includes the derivations of the asymptotic null distribution and the corrected versions of ζ_1 and ζ_2 . Then, the Type I error rate and power of the suggested PFSs are examined for several simulated data sets. In the “Application to Real Data” section, the suggested statistics are computed using a real data set. Conclusions and recommendations are provided in the last section.

This article considers the case of dichotomous items and focuses on IRT-based PFSs.

Method

Review of the ζ_1 and ζ_2 Statistics

Consider a test comprising J dichotomous items that were administered to I examinees. Let us denote the true ability of examinee i as θ_i , and let $\hat{\theta}_i$ denote an estimate of θ_i . Let y_{ij} be the score (i.e., 0 or 1) of examinee i on item j . Let $P_j(\theta_i) = P(y_{ij} = 1)$. For example, for the three-parameter logistic model (3PLM),

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

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

Let us define

$$T(\hat{\theta}_i) = \frac{1}{J} \sum_{j=1}^J P_j(\hat{\theta}_i), \quad G_j = \frac{1}{I} \sum_{i=1}^I P_j(\hat{\theta}_i), \quad \text{and} \quad G = \frac{1}{J} \sum_{j=1}^J G_j = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J P_j(\hat{\theta}_i) = \frac{1}{I} \sum_{i=1}^I T(\hat{\theta}_i). \quad (2)$$

The quantities G_j and G are functions of all the $\hat{\theta}_i$'s in the sample. Therefore, they are not expressed as functions of $\hat{\theta}_i$.

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

$$\zeta_1 = \frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] (G_j - G)}{\sqrt{\sum_{j=1}^J P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)] (G_j - G)^2}}, \quad (3)$$

and the fourth standardized ECI denoted as $ECI4_z$ or ζ_2 for the i th examinee as¹

$$\zeta_2 = \frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] [P_j(\hat{\theta}_i) - T(\hat{\theta}_i)]}{\sqrt{\sum_{j=1}^J P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)] [P_j(\hat{\theta}_i) - T(\hat{\theta}_i)]^2}}. \quad (4)$$

The notations in the above equations are slightly different from those in Tatsuoka (1984) and are used to facilitate the derivations later in this article. Both ζ_1 and ζ_2 , which are based on the caution index (Sato, 1975, 1982), are positive for response patterns with too many correct answers to difficult items or too many incorrect answers to easy items. For example, if the j th item is difficult, then $P_j(\hat{\theta}_i) - T(\hat{\theta}_i) < 0$ for all i , and hence $G_j - G = \frac{1}{J} \sum_{i=1}^I [P_j(\hat{\theta}_i) - T(\hat{\theta}_i)] < 0$. A correct answer on the item by i th examinee means $P_j(\hat{\theta}_i) - y_{ij} < 0$ so that both $(P_j(\hat{\theta}_i) - y_{ij})(G_j - G)$ and $(P_j(\hat{\theta}_i) - y_{ij})[P_j(\hat{\theta}_i) - T(\hat{\theta}_i)]$ are positive. Thus, a large value (e.g., larger than 1.645 at 5% level) of either ζ_1 or ζ_2 indicates person misfit.

The null distributions of PFSs computed using estimated examinee abilities often differ substantially from their theorized null distributions (e.g., Molenaar & Hoijtink, 1990; Snijders, 2001). For example, van Krimpen-Stoop and Meijer (1999) and Snijders (2001) showed that the null distribution of the popular I_z statistic (Drasgow et al., 1985), when computed using $\hat{\theta}_i$, differs from the theorized standard normal null distribution. Snijders suggested a correction of I_z that resulted in the asymptotically standard normal I_z^* statistic.

Tatsuoka (1984) stated, on the basis of results from goodness-of-fit tests of ζ_1 and ζ_2 with the normal distribution in Tatsuoka and Tatsuoka (1982), that there is satisfactory evidence that ζ_1 and ζ_2 computed using estimated abilities may approximately follow the normal distribution under no person misfit. However, she also stated that it would be ideal if the theoretical distribution of these statistics can be derived algebraically. In addition, Li and Olenik (1997) concluded the fit of the standard normal distribution to the null distributions of ζ_1 and ζ_2 to be “less than adequate” for simulated data (p. 228); for example, they found that the variances of the null distributions of these statistics were smaller than 1 (see their Table 1, p. 222).

The ζ_1 and ζ_2 statistics have been used to assess person fit by several researchers such as Birenbaum (1986); Drasgow et al. (1987); Drasgow, Levine, and McLaughlin (1991); Glas and Meijer (2003); Karabatsos (2003); Korir (2014); Li and Olenik (1997); and Molenaar and Hoijtink (1990). Glas and Meijer (2003) found the ζ_2 statistic along with the posterior predictive model checking method to perform the best overall among eight PFSs. Drasgow et al. (1987) found ζ_1 and ζ_2 to have satisfactory power in a comparison of several PFSs. Therefore, the derivation of the theoretical distributions (followed by corrections, if needed, as in Snijders, 2001) of ζ_1 and ζ_2 may be helpful to researchers interested in PFA.

Review of the Asymptotic Corrections of PFSs

Snijders (2001) considered the broader class of standardized PFSs of the form:

$$S(\hat{\theta}_i) = \frac{W(\hat{\theta}_i)}{v(\hat{\theta}_i)}, \quad (5)$$

$$\text{where } W(\hat{\theta}_i) = \sum_{j=1}^J [y_{ij} - P_j(\hat{\theta}_i)] w_j(\hat{\theta}_i), \quad (6)$$

Table 1. Mean and Standard Deviation of the Null Distribution of ζ_1 , ζ_2 , ζ_1^* , and ζ_2^* for the 60-Item Tests.

True θ	<i>M</i>				<i>SD</i>			
	ζ_1	ζ_1^*	ζ_2	ζ_2^*	ζ_1	ζ_1^*	ζ_2	ζ_2^*
−2.0	0.24	0.18	0.14	0.11	0.96	1.01	0.99	1.01
−1.5	0.04	0.00	−0.02	−0.05	0.96	1.00	0.99	1.00
−1.0	−0.07	−0.10	−0.10	−0.12	0.97	1.00	0.98	1.00
−0.5	−0.14	−0.15	−0.16	−0.16	0.98	0.99	0.99	0.99
0.0	−0.15	−0.16	−0.16	−0.17	0.98	0.99	0.99	0.99
0.5	−0.10	−0.12	−0.11	−0.12	0.97	0.99	0.98	1.00
1.0	−0.02	−0.05	−0.03	−0.05	0.95	1.01	0.98	1.01
1.5	0.07	0.02	0.05	0.01	0.90	1.01	0.96	1.01
2.0	0.15	0.08	0.11	0.05	0.87	1.01	0.94	1.01

$w_j(\hat{\theta}_i)$ is a weight function for item j ,

$$\text{and } v^2(\hat{\theta}_i) = \sum_j w_j^2(\hat{\theta}_i) P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)]. \quad (7)$$

Several IRT-based PFSs are special cases of Equation 5. For example, Snijders (2001) showed that the l_z statistic (Drasgow et al., 1985) is a special case of Equation 5 for

$$w_j(\hat{\theta}_i) = \log \frac{P_j(\hat{\theta}_i)}{1 - P_j(\hat{\theta}_i)}. \quad (8)$$

Snijders (2001) proved theoretically that the assumption that $W(\hat{\theta}_i)$ follows a normal distribution with mean 0 and estimated variance $v^2(\hat{\theta}_i)$ is incorrect, which, from Equation 5, means that $S(\hat{\theta}_i)$ does not have an asymptotic standard normal null distribution. Snijders further showed that if $\hat{\theta}_i$ satisfies the condition

$$r_0(\hat{\theta}_i) + \sum_{j=1}^J [y_{ij} - P_j(\hat{\theta}_i)] r_j(\hat{\theta}_i) = 0 \quad (9)$$

for some real-valued $r_0(\hat{\theta}_i)$ and $r_j(\hat{\theta}_i)$,

$$\text{then, asymptotically, } S^*(\hat{\theta}_i) = \frac{W(\hat{\theta}_i) + c_n(\hat{\theta}_i) r_0(\hat{\theta}_i)}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (10)$$

and, equivalently, because $c_n(\hat{\theta}_i) r_0(\hat{\theta}_i) / \psi(\hat{\theta}_i)$ converges to 0 as test length increases,

$$\frac{W(\hat{\theta}_i)}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (11)$$

when there is no person misfit and three regularity conditions are satisfied,² where

$$c_n(\hat{\theta}_i) = \frac{\sum_j P'_j(\hat{\theta}_i) w_j(\hat{\theta}_i)}{\sum_j P'_j(\hat{\theta}_i) r_j(\hat{\theta}_i)}, \quad (12)$$

$$P'_j(\hat{\theta}_i) = \frac{\partial P_j(\hat{\theta}_i)}{\partial \hat{\theta}_i}, \quad (13)$$

$$\tilde{w}_j(\hat{\theta}_i) = w_j(\hat{\theta}_i) - c_n(\hat{\theta}_i) r_j(\hat{\theta}_i), \quad (14)$$

$$\text{and } \psi^2(\hat{\theta}_i) = \sum_j \tilde{w}_j^2(\hat{\theta}_i) P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)]. \quad (15)$$

Snijders (2001) showed (e.g., see his Table 1) that usually,

$$\psi^2(\hat{\theta}_i) < v^2(\hat{\theta}_i). \quad (16)$$

Thus, for long tests and under no person misfit, Equations 5 and 11 imply that the distribution of $S(\hat{\theta}_i)$ can be approximated³ as

$$S(\hat{\theta}_i) = \frac{W(\hat{\theta}_i)}{\psi(\hat{\theta}_i)} \times \frac{\psi(\hat{\theta}_i)}{v(\hat{\theta}_i)} \sim \mathcal{N}\left(0, \frac{\psi^2(\hat{\theta}_i)}{v^2(\hat{\theta}_i)}\right), \quad (17)$$

and Equations 16 and 17 imply that the estimated variance of $S(\hat{\theta}_i)$ is smaller than 1. Therefore, the assumption that $S(\hat{\theta}_i)$ has an asymptotic standard normal null distribution would lead to a less powerful PFA, that is, one would conclude person misfit less often than is ideal.

Comparing Equations 5 and 10, $S^*(\hat{\theta}_i)$ is obtained from the equation for $S(\hat{\theta}_i)$ by adding $c_n(\hat{\theta}_i) r_0(\hat{\theta}_i)$ in the numerator and replacing $v^2(\hat{\theta}_i)$ by $\psi^2(\hat{\theta}_i)$,⁴ in the denominator (Magis, Raiche, & Beland, 2012). Snijders (2001) showed that for the maximum likelihood estimate (MLE), weighted likelihood estimate (WLE; Warm, 1989), and modal a posteriori (MAP) of ability, $r_j(\hat{\theta}_i)$ is given by

$$r_j(\hat{\theta}_i) = \frac{P'_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)]}, \quad (18)$$

whereas Sinharay (2016b) showed that for two robust ability estimates (biweight and Huber; for example, Schuster & Yuan, 2011),

$$r_j(\hat{\theta}_i) = \frac{\omega[a_j(\hat{\theta}_i - b_j)] P'_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i) [1 - P_j(\hat{\theta}_i)]}, \quad (19)$$

where, for the biweight estimate,

$$\omega(t) = \begin{cases} \left[1 - (t/4)^2\right]^2 & \text{if } |t| \leq 4, \\ 0 & \text{otherwise,} \end{cases}$$

and, for the Huber estimate,

$$\omega(t) = \begin{cases} 1 & \text{if } |t| \leq 1, \\ 1/|t| & \text{otherwise} \end{cases}.$$

Snijders (2001) showed that the regularity conditions are satisfied for the common dichotomous IRT models.

Snijders (2001) applied the above asymptotic correction to I_z (Drasgow et al., 1985) to suggest the asymptotically correct I_z^* statistic. Magis, Beland, and Raiche (2014) applied the correction to the infit and outfit mean squares (Wright & Masters, 1982; Wright & Stone, 1979) to suggest corresponding asymptotically correct statistics.

Asymptotic Corrections of the Standardized ECIs

Meijer and Sijtsma (2001) noted that ζ_1 is a special case of Equation 5 for

$$w_j(\hat{\theta}_i) = -(G_j - G). \quad (20)$$

Therefore, the above-mentioned results of Snijders (2001) indicate that if $\hat{\theta}_i$ is the MLE, WLE, MAP, biweight or Huber estimate and $r_j(\hat{\theta}_i)$, $\tilde{w}_j(\hat{\theta}_i)$, $c_n(\hat{\theta}_i)$, and $\psi(\hat{\theta}_i)$ are defined as in Equations 18 or 19, 14, 12, and 15, respectively, and computed using the $w_j(\hat{\theta}_i)$ of Equation 20, then, asymptotically,

$$\zeta_1^* = \frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] (G_j - G) + c_n(\hat{\theta}_i) r_0(\hat{\theta}_i)}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (21)$$

and, equivalently,

$$\frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] (G_j - G)}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (22)$$

when there is no person misfit.

Meijer and Sijtsma (2001) noted that ζ_2 is a special case of Equation 5 for

$$w_j(\hat{\theta}_i) = -[P_j(\hat{\theta}_i) - T(\hat{\theta}_i)]. \quad (23)$$

If $\hat{\theta}_i$ is the MLE, WLE, MAP, biweight or Huber estimate and $r_j(\hat{\theta}_i)$, $\tilde{w}_j(\hat{\theta}_i)$, $c_n(\hat{\theta}_i)$, and $\psi(\hat{\theta}_i)$ are defined as in Equations 18 or 19, 14, 12, and 15, respectively, and computed using the $w_j(\hat{\theta}_i)$ of Equation 23, then, asymptotically,

$$\zeta_2^* = \frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] [P_j(\hat{\theta}_i) - T(\hat{\theta}_i)] + c_n(\hat{\theta}_i) r_0(\hat{\theta}_i)}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (24)$$

and, equivalently,

$$\frac{\sum_{j=1}^J [P_j(\hat{\theta}_i) - y_{ij}] [P_j(\hat{\theta}_i) - T(\hat{\theta}_i)]}{\psi(\hat{\theta}_i)} \sim \mathcal{N}(0, 1), \quad (25)$$

when there is no person misfit.

The above derivations also show that both ζ_1 and ζ_2 have an asymptotic normal null distribution with mean 0 but variance smaller than 1. For example, Equations 3 and 22 imply that for long tests and under no person misfit, the distribution of ζ_1 can be approximated as

$$\zeta_1 = \frac{\sum_{j=1}^J (P_j(\hat{\theta}_i) - y_{ij})(G_j - G)}{\psi(\hat{\theta}_i)} \times \frac{\psi(\hat{\theta}_i)}{v(\hat{\theta}_i)} \sim \mathcal{N}\left(0, \frac{\psi^2(\hat{\theta}_i)}{v^2(\hat{\theta}_i)}\right), \quad (26)$$

and $\psi^2(\hat{\theta}_i)/v^2(\hat{\theta}_i)$ will be shown to be smaller than 1 on average in Table 2. Therefore, the assumption that ζ_1 and ζ_2 have an asymptotic standard normal null distribution would lead to a conservative PFA.

A Simulation Study

The above derivations are based on asymptotic theory and hold only when the number of items on a test is large. Therefore, a detailed simulation study was performed to examine the Type I error rates and power of ζ_1^* and ζ_2^* and to compare them with those of ζ_1 and ζ_2 under a variety of situations including tests that are not long. The simulation study also includes the I_z^* statistic (Snijders, 2001) because it is arguably one of the most popular IRT-based PFSs and is another asymptotically corrected statistic (as are ζ_1^* and ζ_2^*).

Design of the Simulation

The simulation study involved three levels of test length (20 items, 40 items, and 60 items) that represent short, moderate, and long tests.

Scores on the items were generated using the 3PLM. The true slope parameters of all items were generated, as in Glas and Dagohoy (2007), from a log-normal distribution with 0 and 0.25 as the mean and standard deviation, respectively, of the logarithm of the variable. The true difficulty and true guessing parameters of the items were generated from a standard normal distribution and a Uniform(0.05, 0.3) distribution, respectively. The item parameters were estimated (using the marginal maximum likelihood procedure) for each simulated data set before computation of the PFSs.

The Type I error rates of the statistics were computed from data sets that only included score patterns (where a score pattern refers to the collection of scores of one examinee on all the items) that fit the 3PLM. As in Glas and Meijer (2003), the power of the statistics was computed from data sets that included 90% score patterns that fit the 3PLM and 10% score patterns that do not fit the IRT model (or “corrupted”).

Corrupted patterns included score patterns under “cheating” (or “item disclosure”) and “guessing,” as in, for example, Glas and Meijer (2003). When “cheating” was simulated, all corrupted score patterns in a data set represented examinees who cheated on 10% or 30% of the most difficult items. The probability of a correct response to an item on which cheating occurred was set to .8. When “guessing” was simulated, all corrupted score patterns in a data set represented examinees who randomly guessed on 10% or 20% of the easiest items. The probability of a correct response to an item on which an examinee guesses was set to .20. Note that the term “guessing” here is not supposed to denote low-ability examinees choosing random answer options on difficult questions, but is supposed to refer to unmotivated examinees guessing on some items, as in Glas and Meijer; guessing on the easiest items has the most detrimental effect on the estimation of ability (Meijer & Nering, 1997) and is important to detect.

Table 2. Average of the Ratios $\psi(\hat{\theta}_i)/\nu(\hat{\theta}_i)$.

Test length	Statistic	True examinee ability (θ_i)								
		−2.0	−1.5	−1.0	−0.5	0	0.5	1.0	1.5	2.0
20	ζ_1	.97	.98	.99	.99	.99	.97	.94	.90	.87
	ζ_2	.99	.99	.99	.99	.99	.98	.96	.94	.92
40	ζ_1	.96	.97	.98	.99	.99	.96	.93	.88	.85
	ζ_2	.98	.99	.99	.99	.99	.98	.96	.94	.93
60	ζ_1	.95	.96	.98	.99	.99	.97	.94	.90	.86
	ζ_2	.98	.98	.99	.99	.99	.99	.97	.95	.93

Other types of guessing are possible. For example, van Krimpen-Stoop and Meijer (1999) simulated a “random response behavior” under which an examinee guesses on all items. Also note that the score patterns that have been labeled as those representing “cheating” and “guessing” have been referred to as those representing “spuriously high” and “spuriously low” responses, by, for example, Meijer and Nering (1997).

For each simulation condition (where an example of a simulation condition is “20 items and cheating on 10% items”), 1,000 data sets with 1,000 examinees⁵ each were simulated. The true θ_i s of the examinees were randomly drawn from one of the following nine values: −2.0, −1.5, −1.0, −0.5, 0.0, 0.5, 1.0, 1.5, and 2.0. Thus, the 1,000 simulated data sets together under a simulation condition included about 111,111 ($\approx 1,000 \times 1,000/9$) examinees at any specific true θ_i . This design allows the computation of the Type I error rate and power at these θ values accurately. Researchers such as Snijders (2001) found that the Type I error and power of PFSs vary substantially over true θ values. The Type I error rate (at 1% and 5% levels) for a statistic at a specific θ value was computed as the proportion of score patterns simulated with that true θ value in the model-fitting data sets for whom the statistic was statistically significant under a standard normal null distribution assumption. Thus, the Type I error rate at any of the nine values of θ_i is computed from about 111,111 examinees for any test length. The standard error is approximately 0.0003 when the Type I error is close to 0.01, and 0.0007 when the Type I error is close to 0.05.⁶ The power (at a level of 1% or 5%) for a PFS at a specific θ value under, say, cheating on 10% items was computed as the proportion of score patterns representing cheating on 10% items and simulated with that θ value for whom the PFS was statistically significant under a standard normal null distribution assumption.⁷ Thus, the power at each θ value was computed roughly from 11,111 ($\approx 1,000 \times 1,000 \times 0.10/9$) examinees in any simulation condition.

Computation

A computer program written by the author was used for all the computations, including the estimation of item parameters and computation of the ability estimates and the PFSs. The MLE and the WLE were used as the ability estimates and computed using the Newton–Raphson algorithm. The MLEs were restricted to be between −4.5 and 4.5; thus, for example, someone who answered all items correctly⁸ was restricted to have an MLE of 4.5.

The following steps were repeated 1,000 times for any simulation condition:

1. Simulate a set of true item parameters.
2. Simulate 1,000 true ability parameters each of which can be one of nine values (−2.0, −1.5, −1.0, −0.5, 0.0, 0.5, 1.0, 1.5, and 2.0).

3. Use the above true item and ability parameters to simulate what is treated as an “observed” data set with 1,000 examinees. The data set could include only score patterns fitting the 3PLM (for simulation conditions from which Type I error rates are computed) or could include some score patterns fitting the 3PLM and some score patterns representing cheating or guessing (for simulation conditions from which power is computed).
4. Compute the marginal MLEs of the item parameters from the “observed” data set.
5. Compute the ability estimates $\hat{\theta}_i$ (MLE or WLE) of all the examinees from the “observed” data set using the estimated item parameters computed in the previous step.
6. Compute, from the “observed” data set, five PFSs (ζ_1 , ζ_2 , ζ_1^* , ζ_2^* , and I_z^*) for each examinee using $\hat{\theta}_i$ computed in the above step and the estimated item parameters computed in Step 4.

Depending on the simulation condition, the values of the Type I error rates or power for each PFS were computed using the values of the PFS computed in the last step above.

Null Distributions of the PFS

Table 1 shows the mean and standard deviation of the null distribution of ζ_1 , ζ_2 , ζ_1^* , and ζ_2^* for each true ability for the 60-item tests. The table shows that the mean and standard deviation of each of ζ_1^* and ζ_2^* are mostly closer, in comparison with the corresponding uncorrected statistic, to those of a standard normal random variable; this phenomenon is more prominent at the extreme true abilities. For example, at true ability of 2.0, the mean and standard deviation of ζ_1 are 0.15 and 0.87, respectively, but those of ζ_1^* are 0.08 (closer to 0) and 1.01 (closer to 1), respectively; so the asymptotic corrections lead to corrected PFSs whose distributions are closer to the standard normal distribution. Furthermore, the standard deviation of both of ζ_1 and ζ_2 is smaller than 1, which agrees with Table 1 in Li and Olenik (1997). The skewness and kurtosis of ζ_1^* and ζ_2^* are very close to those of ζ_1 and ζ_2 and are not reported here.

Table 2 shows the averages of the ratios $\psi(\hat{\theta}_i)/\nu(\hat{\theta}_i)$ for ζ_1 and ζ_2 for the nine θ values for the different test lengths. The ratios, which are also the standard deviations of ζ_1 and ζ_2 because of Equation 26, are close to 1 for θ values near 0, but become smaller as θ moves away from 0. Furthermore, given Table 2, the standard deviations of ζ_1 and ζ_2 in Table 1 are in agreement with the asymptotic null distributions of ζ_1 and ζ_2 derived in Equations 17 and 26; for example, Table 2 shows that the average of $\psi(\hat{\theta}_i)/\nu(\hat{\theta}_i)$ for ζ_1 for 60-item tests is 0.86 for $\theta = 2.0$; then, Equation 26 suggests that the standard deviation of ζ_1 for $\theta = 2.0$ for 60-item tests should be very close to 0.86 and that is exactly what is observed in Table 1.

Results on Type I Error Rate

The Type I error rates (and power) depend little on whether the MLEs or WLEs were used as the ability estimates. Therefore, only the results for the WLEs will be considered henceforth. Figure 1 shows the Type I error rates of the five statistics for all test lengths for significance levels of 1% and 5%. The title of each panel denotes the test length and the level of significance. In each panel, the true ability is shown along the x -axis, and the Type I error rate is shown along the y -axis. Note that the range of the y -axis differs depending on the significance level. For each statistic, the nine values of the Type I error rate (at true ability of -2.0 , -1.5 , -1.0 , $\dots$, 2.0 , and 2.5) are joined using a line in each panel. A horizontal solid line denotes the significance level in each panel; it is desirable for the Type I error rates to be at or below this line.

Figure 1. Type I error rates for all test lengths at 1% and 5% levels.

Figure 1 shows the following:

- The Type I error rates of ζ_1 are smaller than those of ζ_1^* , and the Type I error rates of ζ_2 are smaller than those of ζ_2^* .
- The Type I error rates of ζ_2^* are smaller than those of ζ_1^* .
- The Type I error rates of l_z^* are often inflated (i.e., exceed the significance level) substantially, especially for large positive abilities and at the 1% level. This phenomenon has been observed by several researchers such as Snijders (2001), Sinharay (2015), and van Krimpen-Stoop and Meijer (1999).
- The Type I error rates of ζ_1^* and ζ_2^* are also inflated in some cases. Compared with l_z^* , the extent of Type I error inflation is smaller for ζ_1^* and smaller still for ζ_2^* . Thus, l_z^* seems to have the most undesirable Type I error rate among the PFSs, and ζ_2^* seems to have the most desirable Type I error rates among the three asymptotically corrected PFSs.

Figure 2. Power at 5% level for 60-item tests.

- The Type I error rates of ζ_1 and ζ_2 are pretty close to the nominal level for some abilities but are smaller than the nominal level at some other abilities such as around zero ability in the three right panels (that indicates that ζ_1 and ζ_2 would have small power at those abilities in the presence of person misfit), especially at the 5% level.
- The extent of inflation of Type I error rates is larger at 1% level compared with 5% level. Larger inflation of Type I error rates of asymptotically corrected PFSs at 1% level than at 5% level has been observed by, for example, Magis et al. (2014), Snijders (2001), and Sinharay (2015).

Results on Power

Figure 2 shows the values of power at 5% level of the PFSs for the different types of person misfit for 60-item tests. Figures A1 to A4 in the online appendix show similar plots for test lengths of 20, 40, and 60 and for 1% and 5% levels.

In each panel of Figure 2, ζ_1^* is more powerful than ζ_1 , and ζ_2^* is more powerful than ζ_2 . In addition,

- Under the two cheating conditions, ζ_1 and ζ_1^* are the most powerful for negative abilities, ζ_2 and ζ_2^* are the most powerful for positive abilities, and I_z^* seems to have the smallest power.

- Under the two guessing conditions, ζ_2 and ζ_2^* are the most powerful for negative abilities but the least powerful for positive abilities.
- For guessing on 10% items, I_z^* seems to have the largest power overall.

Discussion on the Comparative Performance of the Statistics

The above results show the following:

- The suggested asymptotic corrections lead to a benefit at 5% level of significance as the asymptotically corrected PFSs have Type I error rates close to the nominal level and are more powerful than the corresponding uncorrected versions. At 1% level, the asymptotically corrected PFSs are more powerful than the corresponding uncorrected versions but have slightly inflated Type I error rates; the Type I error rates of ζ_2^* at 1% level, though inflated, are satisfactory for $\theta > -2$ according to Cochran's (1952) criterion for robustness that specifies that Type I error rates below 0.06 and 0.015 are satisfactory at levels 0.05 and 0.01, respectively.
- There does not seem to be an overall winner among ζ_1^* , ζ_2^* , and I_z^* , and each of them has some advantages and disadvantages as described below:
  - While I_z^* has the most undesirable Type I error rates, especially at 1% level, it also has the largest power to detect "guessing" on 10% items.
  - The ζ_2^* statistic has the most desirable Type I error rates among the three asymptotically correct PFSs but has the smallest power to detect "guessing" for positive abilities.
  - The ζ_1^* statistic has slightly inflated Type I error rates but has the largest power to detect "cheating" for negative abilities.

The choice of a PFS would depend on the problem at hand. For example, if the investigator is keen to control the Type I error rate, then ζ_2^* is most likely to be preferred because its power is satisfactory overall and its Type I error rate is the closest to the nominal level.

Application to Real Data

The PFSs were computed using data from one form of a licensure examination. The data set was analyzed in several chapters of Cizek and Wollack (in press). The test form includes 170 operational items that are dichotomously scored. Item scores on the form were available for 1,644 examinees. The licensure organization (that administers the examination) identified as compromised 61 items on the form. The organization also flagged 48 individuals on the form as possible cheaters from a variety of statistical analysis and an investigative process that brought in other information. The goal of the PFA here will be to examine whether the PFSs considered here can flag some examinees, especially those who were flagged by the licensure organization; another goal will be to examine whether ζ_1^* and ζ_2^* provide any benefits over ζ_1 , ζ_2 , and I_z^* .

The Rasch model is operationally used in the assessment—The difficulty parameters under the Rasch model were estimated from the data set and were used in the PFA. The WLE was used as the estimate of the examinee ability.

The proportions of examinees flagged by ζ_1 , ζ_1^* , ζ_2 , ζ_2^* , and I_z^* at 5% significance level are provided in Table 3. The first row of numbers provides the proportions flagged among all examinees. The second row of numbers provides the proportions flagged among the 48 examinees who were flagged by the licensure organization; thus, for example, the proportion of 0.15 for ζ_1

Table 3. The Proportions of Examinees Flagged for the Real Data.

Examinees	ζ_1	ζ_1^*	ζ_2	ζ_2^*	I_z^*
All	0.09	0.10	0.09	0.10	0.09
Flagged	0.15	0.21	0.19	0.21	0.17

in the second row indicates that among the 48 examinees flagged by the licensure organization, seven examinees were flagged (note that $7/48 \approx 0.15$) by ζ_1 . Table 3 shows that either of ζ_1^* or ζ_2^* flags more examinees than the corresponding unadjusted version in each row, which agrees with the results from the simulation study.

Table 3 also shows that the proportion flagged for each statistic is about twice among the examinees flagged by the licensure organization compared with among all examinees—This result provides some evidence that the PFSs provide useful information. The larger proportion for either of ζ_1^* or ζ_2^* compared with the corresponding uncorrected version and I_z^* in the second row (i.e., among examinees flagged by the licensure organization) provides evidence that the suggested asymptotic corrections may lead to practical benefits by flagging truly aberrant examinees⁹ more often compared with the uncorrected versions and I_z^* . Further investigation of the results shows that among the examinees flagged by the licensure organization, two were flagged by both ζ_1^* and ζ_2^* but not by I_z^* , and those two and one more were flagged by ζ_1^* but not by ζ_1 . Among these three examinees,

- One examinee had a proportion-correct score of 0.89 on the compromised items but of 0.72 on the noncompromised items; for the examinee, I_z^* , ζ_1 , and ζ_2 were -1.60 , 1.64 , and 1.64 , respectively (thus, none of them was significant at 5% level); however, the examinee's ζ_1^* was 1.77 and ζ_2^* was 1.76 —so both of these were significant at 5% significance level.
- Another examinee had a proportion-correct score of 0.84 on the compromised items but of 0.75 on the compromised item; the examinee was flagged by ζ_1^* , ζ_2 , and ζ_2^* , but not by I_z^* or ζ_1 .

Both of these examinees were also flagged by two statistics for detecting item preknowledge (Sinharay, 2016c).

Conclusion

This article suggests asymptotic corrections of ζ_1 and ζ_2 , two standardized ECIs (Tatsuoka, 1984). The asymptotically corrected versions, ζ_1^* and ζ_2^* , have slightly larger power than the corresponding uncorrected versions. Given the increasing importance of PFA in several contexts such as detection of irregularities in testing behavior (e.g., Tendeiro & Meijer, 2014), the suggested corrections promise to be useful to measurement practitioners.

The asymptotically corrected versions, ζ_1^* and ζ_2^* , were used for PFA in this article. However, they can be used for other purposes as well. For example, Tatsuoka (1984) used ζ_1 and ζ_2 for diagnosing student misconceptions, and ζ_1^* and ζ_2^* can be used for those purposes as well; the latter statistics, which were proved to follow an asymptotic standard normal null distribution, would detect student misconceptions more accurately than the formers.

The ζ_1^* and ζ_2^* statistics are appropriate when an investigator wants to test against an unspecified general alternative (e.g., Glas & Dagohoy, 2007) and may not be the most appropriate PFS for a particular problem. For example, if one needs a PFS for a computerized-adaptive test, a statistic such as that of Sinharay (2016d) will be more appropriate.

Statistical indices for determination of person misfit are useful for providing confirming evidence of inappropriate behavior when evidence from other sources (such as reports from the teachers/proctors in testing centers on examinee behavior) also exists, but the evidence provided by statistical indices is insufficient by itself. For example, Hanson, Harris, and Brennan (1994) commented, in the context of detection of answer copying, that no statistical method on its own can provide conclusive proof that copying occurred; the comment is true about other inappropriate examinee behavior as well. Researchers such as Tendeiro and Meijer (2014) and Meijer, Egberink, Emons, and Sijtsma (2008) recommended complementing PFSs with other sources of information such as seating charts, video surveillance, or follow-up interviews.

There are several limitations of this article and, consequently, several additional topics that can be investigated further. First, the performances of the suggested PFSs were compared with those of a handful of PFSs in this article and can be compared with those of other PFSs such as the H_T statistic that was found quite powerful in several studies, including Karabatsos (2003) and Tendeiro and Meijer (2014). Second, the derivations in this article are based on asymptotic results that hold only when the number of items is large and the IRT model fits the data, and the derivations would not hold in many practical situations such as with short tests; in fact, the Type I error rates of ζ_1^* and ζ_2^* were found to be slightly larger than the nominal level for significance levels around 1% for short or moderately long tests. Thus, it may be safer in practice to obtain the empirical null distribution of ζ_1^* and ζ_2^* by using simulations, as was suggested by van Krimpen-Stoop and Meijer (1999) and Sinharay (2016a), rather than by using the asymptotic results suggested in this article. Third, only the MLE and WLE of ability were used to compute the PFSs in this article. Meijer and Nering (1997) and Sinharay (2016b) found that the use of robust ability estimates led to an increase in power of the I_z or I_z^* statistic, respectively; therefore, the use of robust ability estimates with ζ_1^* and ζ_2^* is a topic of future research. Finally, further research such as Conrad et al. (2010); Conijn, Emons, and Sijtsma (2014); Ferrando (2012); Meijer et al. (2008); and Meijer and Tendeiro (2014) should explore how PFA and the suggested PFSs can be used in practice, for example, in high-stakes educational tests.

Author's Note

Any opinions expressed in this publication are those of the author and not necessarily of Pacific Metrics Corporation.

Acknowledgments

The author would like to thank Hua-Hua Chang, the editor, and the three anonymous reviewers for several helpful comments that led to a significant improvement of the article. The author would also like to thank James A. Wollack for generously sharing a data set that was used in the research that led to this article.

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. The $ECI6_z$ statistic of Tatsuoaka (1984) is identical to $ECI4_z$. So, ζ_2 denotes both $ECI4_z$ and $ECI6_z$.
2. The regularity conditions ensure that $\sqrt{J}(\hat{\theta}_i - \theta_i)$ is bounded, and the item characteristic functions are differentiable twice and their derivatives are bounded.

3. The approximation is facilitated by the result, which follows from derivations in Snijders (2001), that $\psi^2(\hat{\theta}_i)/v^2(\hat{\theta}_i)$ tends to the positive limit $\psi^2(\theta_i)/v^2(\theta_i)$.
4. Or, equivalently, $w_j(\hat{\theta}_i)$ by $\tilde{w}_j(\hat{\theta}_i)$.
5. The Type I error rate and power of the statistics do not depend on the number of examinees in each data set—that is, why this number was not varied and assumed to be 1,000.
6. This is because, for example, when the Type I error is close to 0.01, the corresponding standard error is equal to $\sqrt{0.01 \times 0.99 / 111111} \approx 0.0003$.
7. Note that for each simulation condition representing cheating or guessing, no person-fit statistic (PFS) was computed for the 90% examinees whose score patterns fitted the three-parameter logistic model (3PLM).
8. For whom the likelihood is maximum at $\theta_i = \infty$.
9. It is assumed here that the examinees flagged by the licensure organization are truly aberrant due to the rigorous nature of the flagging process of the organization.

References

- Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement*, 10, 167-174.
- Cizek, G. J., & Wollack, J. A. (in press). *Handbook of detecting cheating on tests*. Washington, DC: Routledge.
- Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23, 315-345.
- Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic l_z -based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, 38, 122-136.
- Conrad, K. J., Bezruczko, N., Chan, Y., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence*, 106, 92-100.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59-79.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, 15, 171-191.
- 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.
- Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, 52, 718-722.
- Ferrando, P. J. (2015). Assessing person-fit in typical-response measures. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 128-156). New York, NY: Routledge.
- Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, 72, 159-180.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27, 217-233.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1994). *A comparison of several statistical methods for examining allegations of copying* (ACT Research Report Series No. 87-15). Iowa City, IA: American College Testing.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298.
- Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, 56, 213-228.
- Korir, D. K. (2014). The effects of item difficulty and examinees ability on the effectiveness of ECIZ4 appropriateness index. *Education*, 4(4), 88-94.
- Li, M. F., & Olenik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21, 215-231.

- Magis, D., Beland, S., & Raiche, G. (2014). Snijders's correction of infit and outfit indexes with estimated ability level: An analysis with the Rasch model. *Journal of Applied Measurement*, 15, 82-93.
- Magis, D., Raiche, G., & Beland, 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., Egberink, I. J., Emons, W. H., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, 90, 227-238.
- Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, 21, 321-336.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Meijer, R. R., & Tendeiro, J. N. (2014). *LSAC Research Report Series: The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us*. Newtown, PA: Law School Admission Council.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.
- Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test securities irregularities*. Washington DC: Council of Chief State School Officers.
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meiji Tosho.
- Sato, T. (1982). *Application of S-P curve theory*. Tokyo, Japan: Meiji Tosho.
- Schuster, C., & Yuan, K. H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, 36, 720-735.
- Sinharay, S. (2015). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9465-x
- Sinharay, S. (2016a). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement*, 53, 63-85.
- Sinharay, S. (2016b). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of Mathematical and Statistical Psychology*, 69, 175-193.
- Sinharay, S. (2016c). *Detection of item preknowledge using likelihood ratio test and score test* (Manuscript Under Review).
- Sinharay, S. (2016d). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi:10.3102/1076998616658331
- Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, 359-372.
- Snijders, T. (2001). Asymptotic 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. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, 51, 239-259.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23, 327-345.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis* [Computer software]. Chicago, IL: Mesa Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.