Article

The Standardized S-X^2 Statistic for Assessing Item Fit

Applied Psychological Measurement
2023, Vol. 47(1) 3–18

DOI: 10.1177/01466216221108077
journals.sagepub.com/home/apm
SAGE

Zhuangzhuang Han[1], Sandip Sinharay[1], Matthew S. Johnson[1], and Xiang Liu[1]

Abstract

The S-X^2 statistic (Orlando & Thissen, 2000) is popular among researchers and practitioners who are interested in the assessment of item fit. However, the statistic suffers from the Chernoff–Lehmann problem (Chernoff & Lehmann, 1954) and hence does not have a known asymptotic null distribution. This paper suggests a modified version of the S-X^2 statistic that is based on the modified Rao–Robson χ^2 statistic (Rao & Robson, 1974). A simulation study and a real data analyses demonstrate that the use of the modified statistic instead of the S-X^2 statistic would lead to fewer items being flagged for misfit.

Keywords

item response theory model fit, Orlando-Thissen statistic, Pearson's, statistic, Rao-Robson's modified, statistic

Introduction

The Standard 4.10 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) recommends documenting evidence of model-data fit when an item response theory (IRT) model is employed in test development and score reporting. In practice, analysis of model-data fit for IRT models involves the use of item-fit residuals and χ^2-type statistics (Hambleton & Han, 2005). Among the χ^2-type statistics for IRT models, the S-X^2 statistic (Orlando & Thissen, 2000) is popular, presumably because of four reasons. First, to compute S-X^2, one has to divide the examinees into groups based on their observed total scores rather than the estimated abilities. Second, S-X^2 has been found to perform respectably in terms of Type I error rates and power in simulation studies (e.g., Glas & Suarez-Falcón, 2003; Sinharay, 2006; Sinharay & Lu, 2008; Stone & Zhang, 2003). Third, the simple and intuitive nature of S-X^2 has allowed it to be easily generalized to cases with polytomous items (Kang & Chen, 2008, 2010),

[1]Educational Testing Service, Princeton NJ, USA

Corresponding Author:
Sandip Sinharay, Department of Research and Development, Educational Testing Service, MS 12T, 660 Rosedale Road, Princeton NJ 08541, USA.
Email: ssinharay@ets.org

multidimensional examinee abilities (Zhang & Stone, 2007), unfolding models (Roberts, 2008), and cognitive diagnostic models (e.g., Sorrel et al., 2017). Fourth, S-X^2 is implemented in multiple IRT software packages including irtplay (Lim, 2020), mirt (Chalmers, 2012), and IRTPRO (Cai et al., 2011)

Notwithstanding these appealing features, S-X^2 should not be used without considering its limitations. As noted by researchers such as Sinharay (2006), S-X^2, which is a special case of the Pearson's χ^2 statistic (Pearson, 1900), does not have a known asymptotic null distribution in typical IRT applications where the traditional marginal maximum likelihood estimates (MMLEs) of item parameters are used to compute the statistic. Instead, the values of S-X^2 are stochastically larger than those from the theorized (χ^2) distribution of the statistic. As a consequence, the Type I error rates of S-X^2 tend to be slightly larger than the nominal level even for large samples, which has been observed in multiple simulation studies (e.g., Glas & Suarez-Falcón, 2003; Sinharay, 2006; Sinharay & Lu, 2008). The aim of this paper is to introduce a modified S-X^2 statistic that has a known χ^2 asymptotic null distribution.

The next section includes a review of the Pearson's χ^2 statistic used for assessing general model-data fit and the S-X^2 statistic (Orlando & Thissen, 2000) for assessing item fit, followed by a brief review of a potential problem associated with the use of the Pearson's χ^2 statistic (Chernoff & Lehmann, 1954). The section also includes a description of the modified Pearson's χ^2 statistic that Rao and Robson (1974) suggested to overcome the Chernoff–Lehmann problem. The method section presents the details of our modified S-X^2 statistic that is a special case of the modified Pearson's χ^2 statistic. The section on simulation studies compares the modified S-X^2 statistic with the original S-X^2 statistic with respect to Type I error rates and power. The two statistics are compared using a real data set in the penultimate section. Conclusions and recommendations are provided in the last section. Although the S-X^2 statistic has been extended to tests with polytomously scored items (Kang & Chen, 2008, 2010), we will only consider tests with dichotomously scored items.

Background: Pearson's χ^2, Orlando-Thissen's S-X^2, Chernoff–Lehmann Problem, and Rao–Robson's Modified χ^2

Pearson's χ^2 Statistic

Let us assume that a sample with N independent observations, $y_1, y_2, \ldots, y_N$, is available from a population. Suppose that $p(y_i; \boldsymbol{\eta})$, the probability distribution of y_i, involves a parameter vector $\boldsymbol{\eta}$ with L elements. Suppose that the observations are partitioned into K groups (or cells) and the proportion of observations belonging to group k is $p_k = \frac{N_k}{N}$, where N_k represents the number of observations in group k, $k = 1, 2, \ldots, K$. Let $\pi_k(\boldsymbol{\eta})$ denote the expected value of p_k under the assumed probability distribution.

Pearson's χ^2 statistic (Pearson, 1900) for assessing goodness of fit, denoted henceforth as P-X^2, is defined as

$$\mathrm{P}-X^2 = N\sum_{k=1}^{K}\frac{(p_k-\pi_k(\boldsymbol{\eta}))^2}{\pi_k(\boldsymbol{\eta})} = [\boldsymbol{u}(\boldsymbol{\eta})]^\top \boldsymbol{u}(\boldsymbol{\eta}), \tag{1}$$

where

$$\boldsymbol{u}(\boldsymbol{\eta}) = \sqrt{N}\left(\frac{p_1-\pi_1(\boldsymbol{\eta})}{\sqrt{\pi_1(\boldsymbol{\eta})}}, \frac{p_2-\pi_2(\boldsymbol{\eta})}{\sqrt{\pi_2(\boldsymbol{\eta})}}, \ldots, \frac{p_K-\pi_K(\boldsymbol{\eta})}{\sqrt{\pi_K(\boldsymbol{\eta})}}\right)^\top. \tag{2}$$

In practice, the parameter vector $\boldsymbol{\eta}$ is unknown and P-X^2 is computed by replacing $\boldsymbol{\eta}$ by $\hat{\boldsymbol{\eta}}$, which is the maximum likelihood estimate (MLE) of $\boldsymbol{\eta}$, and is assumed to follow a χ^2 distribution with

$K - L - 1$ degrees of freedom (df), or, the χ^2_{K-L-1} distribution, for large samples under no item misfit.

Orlando and Thissen's S-X^2 Statistic

Orlando and Thissen (2000) developed the S-X^2 statistic, which is a special case of the Pearson's χ^2 statistic, to assess item fit in the context of IRT models for dichotomously scored items. Suppose that we are interested in assessing item fit for a J-item test. To compute S-X^2 for a given *item of interest*, the examinees are divided into $(J + 1)$ groups, where group k includes all the examinees whose raw score is k. Let N_k denote the size of group k. One then computes, for each group k, O_k, which is the observed proportion of test-takers in the group who answered the item correctly. The statistic S-X^2 for the item is then computed as

$$\mathrm{S}-X^2 = \sum_{k=1}^{K} \frac{N_k[O_k - E_k(\boldsymbol{\eta})]^2}{E_k(\boldsymbol{\eta})[1 - E_k(\boldsymbol{\eta})]} = [\boldsymbol{v}(\boldsymbol{\eta})]^\top \boldsymbol{v}(\boldsymbol{\eta}), \tag{3}$$

where $K = J - 1$, $E_k(\boldsymbol{\eta})$ is the expected value, under the IRT model, of O_k

$$\boldsymbol{v}(\boldsymbol{\eta}) = \left(\frac{\sqrt{N_1}[O_1 - E_1(\boldsymbol{\eta})]}{\sqrt{E_1(\boldsymbol{\eta})[1 - E_1(\boldsymbol{\eta})]}}, \frac{\sqrt{N_2}[O_2 - E_2(\boldsymbol{\eta})]}{\sqrt{E_2(\boldsymbol{\eta})[1 - E_2(\boldsymbol{\eta})]}}, \ldots, \frac{\sqrt{N_K}[O_K - E_K(\boldsymbol{\eta})]}{\sqrt{E_K(\boldsymbol{\eta})[1 - E_K(\boldsymbol{\eta})]}} \right)^\top, \tag{4}$$

and the $L \times 1$ vector $\boldsymbol{\eta}$ includes the parameters of the item of interest, that is, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$, where L could vary over the items depending on the assumed IRT model, and, for example, would be equal to 2 if the two-parameter logistic (2PL) model is used. Let $v_k(\boldsymbol{\eta})$ denote the k-th element of $\boldsymbol{v}(\boldsymbol{\eta})$.

In computing the S-X^2 statistic, the number of examinee groups (K) is typically equal to $J - 1$ because $O_0 = E_0(\boldsymbol{\eta}) = 0$ and $O_J = E_J(\boldsymbol{\eta}) = 1$ for any data set. For small samples, to ensure that the expected number of examinees is not too small in any examinee group, some groups may be merged and K can be set equal to a number smaller than $J - 1$. In the simulations and empirical data examples for this paper, groups with fewer than 5 expected number of test-takers were merged, as was recommended by Orlando and Thissen (2000). However, for the sake of simplicity, merging is not considered in the theoretical derivations.

The expected proportion of examinees for group k, $E_k(\boldsymbol{\eta})$, is computed as

$$E_k(\boldsymbol{\eta}) = \frac{\int P(Y = 1|\theta, \boldsymbol{\eta}) P(T_{-1} = k - 1|\theta, \boldsymbol{\eta}) \psi(\theta) \, d\theta}{\int P(T = k|\theta, \boldsymbol{\eta}) \psi(\theta) \, d\theta}, \tag{5}$$

where Y is the score of a randomly chosen examinee on the item of interest, $P(Y = 1|\theta, \boldsymbol{\eta})$ is the probability that Y is equal to 1 given examinee ability θ and item parameters $\boldsymbol{\eta}$, T is the total (raw) score on the test, T_{-1} is the *rest score*, or the total score on all items except the item of interest, $P(T = k|\theta, \boldsymbol{\eta})$ is the probability that T is equal to k given ability θ and item parameters $\boldsymbol{\eta}$, $P(T_{-1} = k - 1|\theta, \boldsymbol{\eta})$ is the probability that the rest score given ability θ is equal to $k - 1$, and $\psi(\theta)$ is the population distribution of the examinee ability and typically assumed to be the standard normal distribution. The integrals in equation (5) are approximated using numerical integration.

The expressions $P(Y = 1|\theta, \boldsymbol{\eta})$, $P(T_{-1} = k - 1|\theta, \boldsymbol{\eta})$, and $P(T = k|\theta, \boldsymbol{\eta})$ depend on the IRT model fitted to the data. If, for example, the 2PL model is used, then

$$P(Y = 1|\theta, \boldsymbol{\eta}) = \frac{\exp[a(\theta - b)]}{1 + \exp[a(\theta - b)]},$$

where a and b, respectively, are the slope and difficulty parameters of the item of interest. Also, the terms $P(T_{-1} = k - 1|\theta, \boldsymbol{\eta})$ and $P(T = k|\theta, \boldsymbol{\eta})$ are computed using the Lord–Wingersky recursion formula (Lord & Wingersky, 1984).

Orlando and Thissen (2000) assumed that the asymptotic null distribution of S-X^2 is the χ^2_{K-L} distribution.

The Chernoff–Lehmann Problem with the Pearson's χ^2 Statistic

A critical step in defining P-X^2, the Pearson's χ^2 Statistic, is the partitioning of the data into K groups. Under the setup of subsection 2.1, the grouped data comprise $O_k = Np_k$, $k = 1, 2, \ldots, K$,. Because the O_k's follow the multinomial distribution (e.g., Agresti, 2013, p. 6), the log-likelihood of $\boldsymbol{\eta}$ based on the grouped data is given by

$$\log \prod_k [\pi_k(\boldsymbol{\eta})]^{Np_k} = N \sum_k p_k \log \pi_k(\boldsymbol{\eta}). \tag{6}$$

Fisher (1924) proved that if P-X^2 is computed using the estimated parameter vectors $\tilde{\boldsymbol{\eta}}$ that maximizes the log-likelihood provided in equation (6), then the asymptotic null distribution of P-X^2 is the χ^2_{K-L-1} distribution. That is, for large samples and under no model misfit

$$\mathrm{P} - X^2 = \left[\boldsymbol{u}\left(\tilde{\boldsymbol{\eta}}\right)\right]^\top \boldsymbol{u}\left(\tilde{\boldsymbol{\eta}}\right) \sim \chi^2_{K-L-1}. \tag{7}$$

The distribution reflects a loss of 1 df for each parameter that is estimated. The estimate $\tilde{\boldsymbol{\eta}}$ is often referred to as the minimum χ^2 estimator (e.g., Harris & Kanji, 1983).

Let $\hat{\boldsymbol{\eta}}$ denote the MLE of $\boldsymbol{\eta}$, which is computed by maximizing

$$\sum_{i=1}^{N} \log f(\boldsymbol{y}_i, \boldsymbol{\eta}),$$

which is the log-likelihood for the original/ungrouped data.

Chernoff and Lehmann (1954) proved that if one uses $\hat{\boldsymbol{\eta}}$ to compute P-X^2, the corresponding statistic

$$\mathrm{P} - X^2 = \left[\boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right)\right]^\top \boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right) \sim \chi^2_{K-L-1} + \sum_{l=1}^{L} \lambda_l\left(\hat{\boldsymbol{\eta}}\right)\chi^2_1, \tag{8}$$

where $0 < \lambda_l(\boldsymbol{\eta}) < 1$; that is, the statistic is somewhere between a χ^2_{K-L-1} variable and a χ^2_{K-1} variable on average. Equation (8) implies that if a statistic of the form $\left[\boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right)\right]^\top \boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right)$ is used to assess item fit and the χ^2_{K-L-1} distribution is used to approximate the limiting distribution of the statistic, the null hypothesis of adequate model fit will be rejected more often than is appropriate, which would result in an inflated Type I error rate of the fit-assessment approach.

Equations (1) and (4) imply that the S-X^2 statistic is a special case of the Pearson's χ^2 statistic. In addition, S-X^2 is computed using the MMLE of the item parameters based on the original/ungrouped data and yet is assumed to have a χ^2_{J-L-1} asymptotic null distribution (Orlando & Thissen, 2000). Such a use of S-X^2 is exactly like the use of the Pearson's χ^2 statistic along with the χ^2_{K-L-1} asymptotic null distribution. Therefore, S-X^2 is expected to suffer from the Chernoff-Lehmann Problem and is expected to follow not a χ^2 distribution, but a distribution like

the one given by equation (8). Thus, S-X^2 is expected to be larger on average than a χ^2_{J-L-1} random variable for large samples under no model misfit. Existing simulation studies that examined the Type I error rates of S-X^2 corroborate this fact. Glas and Suarez-Falcón (2003), Sinharay (2006), and Sinharay and Lu (2008) found in simulation studies that the Type I error rates of S-X^2 are slightly inflated when it is computed using the MMLEs of item parameters from ungrouped data and is assumed to have the χ^2_{J-L-1} asymptotic null distribution. For example, Table 1 of Glas and Suarez-Falcón (2003) shows that the Type I error rates of S-X^2 at 5% significance level are 0.08, 0.08, and 0.07, respectively, for sample sizes 500, 1,000, and 4000 for 10-item tests. The resampling-based approaches developed by Sinharay (2006), Stone (2000), Stone and Zhang (2003), which involve the determination of the null distribution of S-X^2 using simulations, offer alternative solutions and successfully avoid the use of an inaccurate asymptotic null distribution, but these approaches are computation-intensive. The use of the minimum χ^2 estimator $\tilde{\boldsymbol{\eta}}$ and the P-X^2 statistics defined in equation (7) is another possible approach to attain the target Type I error rate. However, $\widehat{\boldsymbol{\eta}}$ is a more efficient estimator compared to $\tilde{\boldsymbol{\eta}}$ because the former utilizes more information than the latter (e.g., Rao, 1962; Rao & Robson, 1974). Also, $\hat{\boldsymbol{\eta}}$ is more popular than $\tilde{\boldsymbol{\eta}}$. For example, the former is implemented in several publicly available IRT software packages such as BILOG (Mislevy & Bock, 1991), MULTILOG (Thissen, 1991), and PARSCALE (Muraki & Bock, 2003). Further, a χ^2-type statistic that utilizes $\hat{\boldsymbol{\eta}}$ rather than $\tilde{\boldsymbol{\eta}}$ is likely to be more useful and popular among researchers and practitioners.

The Modified χ^2 Statistic of Rao and Robson

One solution to the abovementioned Chernoff–Lehmann problem is to modify P-X^2 in a way such that the modified statistic has a known asymptotic null distribution.

One modification of the Pearson's χ^2 statistic was suggested by Rao and Robson (1974) and is computed as

$$\mathrm{P}-X^2_{RR} = \left[\boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right)\right]^{\top} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{u}\left(\widehat{\boldsymbol{\eta}}\right)} \boldsymbol{u}\left(\hat{\boldsymbol{\eta}}\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{u}(\hat{\boldsymbol{\eta}})}$ is the approximate covariance matrix of $\boldsymbol{u}(\widehat{\boldsymbol{\eta}})$ for large samples. The modification is essentially a standardization of $\boldsymbol{u}(\hat{\boldsymbol{\eta}})$ such that $\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{u}(\hat{\boldsymbol{\eta}})}\boldsymbol{u}(\hat{\boldsymbol{\eta}})$ follows a multivariate normal distribution for large samples under no model misfit, and, consequently

$$\mathrm{P}-X^2_{RR} \sim \chi^2_{K-1}.$$

Note that there is no loss of df for parameter estimation in the null distribution of the $\mathrm{P}-X^2_{RR}$ statistic. Rao and Robson (1974) found that $\mathrm{P}-X^2_{RR}$ has larger power than the Pearson's χ^2

Table 1. The Type I Error Rates of S-X^2 and $S-X^2_{RR}$ for the 2PL model.

Test		Sample size			
Length	Statistic	500	1000	2000	4000
10	S-X^2	0.092	0.087	0.074	0.068
	$S-X^2_{RR}$	0.035	0.042	0.043	0.044
20	S-X^2	0.072	0.067	0.061	0.057
	$S-X^2_{RR}$	0.054	0.048	0.041	0.040
40	S-X^2	0.062	0.057	0.053	0.051
	$S-X^2_{RR}$	0.054	0.053	0.049	0.047

statistic computed using the minimum χ^2 estimator defined in equation (7)—this result is presumably due to the larger degrees of freedom of the former statistic compared to the latter statistic.

In this paper, we borrow the idea underlying $\mathrm{P}-X^2_{RR}$ and derive the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$. The matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$ allows us to compute the statistic $\mathrm{S}-X^2_{RR}$, which is a special case of the $\mathrm{P}-X^2_{RR}$ statistic and is a modified version of the S-X^2 statistic, as

$$\mathrm{S}-X^2_{RR}=\left[\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right)\right]^{\top}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right). \tag{9}$$

Further

$$\mathrm{S}-X^2_{RR}\sim\chi^2_{J-1}$$

(Rao & Robson, 1974). The key of this modification is the computation of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$. The detailed derivation of the matrix is provided below.

Method: Derivation of the Covariance Matrix Required in $\mathrm{S}-X^2_{RR}$

To obtain $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$, we first approximate $\boldsymbol{v}(\hat{\boldsymbol{\eta}})$ using the first-order Taylor series expansion (e.g., Lehmann & Casella, 1998, p. 77) around $\boldsymbol{\eta_0}$ as

$$\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right)\approx\boldsymbol{v}(\boldsymbol{\eta_0})+\boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta_0}\right), \tag{10}$$

where $\boldsymbol{\eta_0}$ is the unknown true item parameter vector

$$\boldsymbol{v}(\boldsymbol{\eta_0})=\left(\frac{\sqrt{N_1}[O_1-E_1(\boldsymbol{\eta_0})]}{\sqrt{E_1(\boldsymbol{\eta_0})[1-E_1(\boldsymbol{\eta_0})]}},\frac{\sqrt{N_2}[O_2-E_2(\boldsymbol{\eta_0})]}{\sqrt{E_2(\boldsymbol{\eta_0})[1-E_2(\boldsymbol{\eta_0})]}},\ldots,\frac{\sqrt{N_K}[O_K-E_K(\boldsymbol{\eta_0})]}{\sqrt{E_K(\boldsymbol{\eta_0})[1-E_K(\boldsymbol{\eta_0})]}}\right)^{\top}, \tag{11}$$

and $\boldsymbol{A_0}$ is a $K\times L$ matrix whose (k, l)-th element is given by

$$\begin{aligned}(\boldsymbol{A_0})_{k,l}&=\frac{\partial v_k(\boldsymbol{\eta})}{\partial E_k(\boldsymbol{\eta})}\frac{\partial E_k(\boldsymbol{\eta})}{\partial\eta_l}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta_0}}\\&=N_k^{1/2}\left[-\frac{1}{E_k(\boldsymbol{\eta_0})^{1/2}(1-E_k(\boldsymbol{\eta_0}))^{1/2}}+\frac{(E_k(\boldsymbol{\eta_0})-0.5)(O_k-E_k(\boldsymbol{\eta_0}))}{E_k(\boldsymbol{\eta_0})^{3/2}(1-E_k(\boldsymbol{\eta_0}))^{3/2}}\right]\frac{\partial E_k(\boldsymbol{\eta})}{\partial\eta_l}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta_0}}.\end{aligned} \tag{12}$$

Note that for large values of N_k, O_k is approximately equal to $E_k(\boldsymbol{\eta_0})$, and, consequently, $(\boldsymbol{A_0})_{k,l}$ can be approximated as

$$(\boldsymbol{A_0})_{k,l}\approx-\sqrt{\frac{N_k}{E_k(\boldsymbol{\eta_0})(1-E_k(\boldsymbol{\eta_0}))}}\;\frac{\partial E_k(\boldsymbol{\eta})}{\partial\eta_l}\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta_0}}. \tag{13}$$

Equation (10) implies that

$$\boldsymbol{\Sigma}_{\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right)}\approx\boldsymbol{\Sigma}_{\boldsymbol{v}(\boldsymbol{\eta_0})}+2\mathrm{Cov}\left[\boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta_0}\right),\boldsymbol{v}(\boldsymbol{\eta_0})\right]+\boldsymbol{A_0}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{A}_{\boldsymbol{0}}^{\top}. \tag{14}$$

Among the terms in equation (14), the elements of $\boldsymbol{A_0}$ can be approximated using equation (13) and $\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}$, which is the variance-covariance matrix among the estimates of the item parameters, can be obtained from the IRT software that was used to fit the IRT model to the data set.[1] The computation of the other terms, $\boldsymbol{\Sigma}_{\boldsymbol{v}(\boldsymbol{\eta_0})}$ and $\mathrm{Cov}[\boldsymbol{A_0}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta_0}),\boldsymbol{v}(\boldsymbol{\eta_0})]$, are described below.

Computation of $\Sigma_{\mathbf{v}(\boldsymbol{\eta}_0)}$

Because of equation (11), the diagonal elements of $\Sigma_{\boldsymbol{v}(\boldsymbol{\eta}_0)}$ are terms such as $\mathrm{Var}(v_k(\boldsymbol{\eta}_0))$, where

$$v_k(\boldsymbol{\eta}_0) = \frac{\sqrt{N_k}[O_k - E_k(\boldsymbol{\eta}_0)]}{\sqrt{E_k(\boldsymbol{\eta}_0)[1 - E_k(\boldsymbol{\eta}_0)]}}, k = 1, 2, \ldots, K,$$

computed at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and the off-diagonal elements of $\Sigma_{\boldsymbol{v}(\boldsymbol{\eta}_0)}$ are terms such as $\mathrm{Cov}(v_{k_1}(\boldsymbol{\eta}_0), v_{k_2}(\boldsymbol{\eta}_0))$ for $k_1 \neq k_2 = 1, 2, \ldots, K$ computed at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$.

Because the variance of O_k computed at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ is $E_k(\boldsymbol{\eta}_0)[1 - E_k(\boldsymbol{\eta}_0)]/N_k$, $v_k(\boldsymbol{\eta}_0)$ is standardized, that is, its variance is 1 for $k = 1, 2, \ldots, K$. So, the diagonal elements of $\Sigma_{\boldsymbol{v}(\boldsymbol{\eta}_0)}$ are all equal to 1. Because the quantities $E_{k_1}(\boldsymbol{\eta}_0)$ are constants, $\mathrm{Cov}(v_{k_1}(\boldsymbol{\eta}_0), v_{k_2}(\boldsymbol{\eta}_0))$ is a multiple of $\mathrm{Cov}(O_{k_1}, O_{k_2})$, the covariance of O_{k_1} and O_{k_2}, computed at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$. Appendix A includes a proof that $\mathrm{Cov}(O_{k_1}, O_{k_2})$, computed at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, is approximately equal to 0 for large samples. Therefore, the off-diagonal elements of $\Sigma_{\boldsymbol{v}(\boldsymbol{\eta}_0)}$ are all approximately equal to 0 for large samples.

Consequently, for large samples

$$\Sigma_{\boldsymbol{v}(\boldsymbol{\eta}_0)} \approx \boldsymbol{I}_{\boldsymbol{K}}, \tag{15}$$

where $\boldsymbol{I}_{\boldsymbol{K}}$ denotes an identity matrix of dimension $K \times K$.

Computation of $\mathrm{Cov}[\boldsymbol{A}_0(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0), \mathbf{v}(\boldsymbol{\eta}_0)]$

The grouped data in the context of item-fit analysis comprise the quantities $N_k O_k$ and $N_k(1 - O_k)$, which are the numbers of correct and incorrect answers on the item of interest for examinee group k. The log-likelihood of these grouped data is provided by

$$\tilde{\ell}(\hat{\boldsymbol{\eta}}) = \log \prod_k E_k(\hat{\boldsymbol{\eta}})^{N_k O_k} (1 - E_k(\hat{\boldsymbol{\eta}}))^{N_k(1-O_k)} = \sum_k [N_k O_k \log(E_k(\hat{\boldsymbol{\eta}})) + N_k(1 - O_k) \log(1 - E_k(\hat{\boldsymbol{\eta}}))].$$

As mentioned earlier, the minimum χ^2 estimator $\tilde{\boldsymbol{\eta}}$ is obtained by solving

$$\frac{\partial \tilde{\ell}(\boldsymbol{\eta})}{\partial \eta_l} = \sum_k \left[\frac{N_k O_k}{E_k(\boldsymbol{\eta})} - \frac{N_k(1 - O_k)}{1 - E_k(\boldsymbol{\eta})} \right] \frac{\partial E_k(\boldsymbol{\eta})}{\partial \eta_l} = 0, l = 1, 2, \ldots, L, \tag{16}$$

or by solving

$$\sum_k \frac{N_k[O_k - E_k(\boldsymbol{\eta})]}{E_k(\boldsymbol{\eta})[1 - E_k(\boldsymbol{\eta})]} \frac{\partial E_k(\boldsymbol{\eta})}{\partial \eta_l} = 0, l = 1, 2, \ldots, L.$$

Therefore, the solution $\tilde{\boldsymbol{\eta}}$ to the above equations satisfies

$$\frac{\partial \tilde{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}} = \mathbf{0}_{L \times 1}, \tag{17}$$

where $\mathbf{0}_{L \times 1}$ is a vector of length L whose elements are zeroes. Also note that Equations (11), (13), and (16) imply that

$$\frac{\partial \tilde{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta} = \boldsymbol{\eta}_0} = -\boldsymbol{A}_0^{\top} \boldsymbol{v}(\boldsymbol{\eta}_0). \tag{18}$$

By applying the Taylor series expansion around $\boldsymbol{\eta} = \boldsymbol{\eta_0}$ to $\frac{\partial \tilde{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}}$ and using the result provided in Equations (17) and (18), we obtain

$$\frac{\partial \tilde{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}} = \mathbf{0}_{L\times 1} \approx -\boldsymbol{A_0}^{\top}\boldsymbol{v}(\boldsymbol{\eta_0}) + \boldsymbol{B_0}\left(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta_0}\right), \tag{19}$$

where

$$\boldsymbol{B_0} = \frac{\partial^2 \tilde{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta_0}}.$$

Equation (19) implies that

$$\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top}\mathbf{v}(\boldsymbol{\eta_0}) - \tilde{\boldsymbol{\eta}} + \boldsymbol{\eta_0} \approx 0$$

or

$$\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top}\boldsymbol{v}(\boldsymbol{\eta_0}) + \hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}} \approx \hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}. \tag{20}$$

Using equation (20), we can express the covariance $\mathrm{Cov}[\boldsymbol{A_0}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \boldsymbol{v}(\boldsymbol{\eta_0})]$ in equation (14) as

$$\begin{aligned}\mathrm{Cov}\left[\boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}\right), \boldsymbol{v}(\boldsymbol{\eta_0})\right] &\approx \mathrm{Cov}\left[\boldsymbol{A_0}\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top}\boldsymbol{v}(\boldsymbol{\eta_0}) + \boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}\right), \boldsymbol{v}(\boldsymbol{\eta_0})\right] \\ &= \boldsymbol{A_0}\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{v}(\boldsymbol{\eta_0})} + \mathrm{Cov}\left[\boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}\right), \boldsymbol{v}(\boldsymbol{\eta_0})\right].\end{aligned} \tag{21}$$

However, note that $\mathrm{Cov}[\boldsymbol{A_0}(\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}), \boldsymbol{v}(\boldsymbol{\eta_0})]$, the second term in the right side of equation (21), converges to a matrix of zeroes since $\boldsymbol{A_0}$ is a matrix of constants and $\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}$, which is the difference between two sets of item parameter estimates, converges to a zero vector as sample size increases. Therefore, equation (21) yields the result that

$$\mathrm{Cov}\left[\boldsymbol{A_0}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}\right), \boldsymbol{v}(\boldsymbol{\eta_0})\right] = \boldsymbol{A_0}\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{v}(\boldsymbol{\eta_0})} \tag{22}$$

Equations (14), (15), and (22) imply that

$$\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})} \approx \boldsymbol{I_K} + 2\boldsymbol{A_0}\boldsymbol{B_0}^{-1}\boldsymbol{A_0}^{\top} + \boldsymbol{A_0}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{A_0}^{\top}. \tag{23}$$

Although the minimum χ^2 estimator appears in the above derivation, one does not have to compute the estimator to compute $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$. That is because $\boldsymbol{A_0}$ and $\boldsymbol{B_0}$ can be adequately approximated using the MLE $\hat{\boldsymbol{\eta}}$ that is an accurate estimator of $\boldsymbol{\eta_0}$ for common IRT models (e.g., Harwell et al., 1988).

After approximating $\boldsymbol{\Sigma}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}$ using equation (23), one can compute our modified version of S-X^2 as

$$\mathrm{S} - X^2_{RR} = \left[\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right)\right]^{\top}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{v}(\hat{\boldsymbol{\eta}})}\boldsymbol{v}\left(\hat{\boldsymbol{\eta}}\right), \tag{24}$$

where $\boldsymbol{v}(\hat{\boldsymbol{\eta}})$ is computed using equation (4) after replacing $\boldsymbol{\eta}$ by $\hat{\boldsymbol{\eta}}$. The asymptotic null distribution of $\mathrm{S} - X^2_{RR}$ is a χ^2_{J-1} distribution (Rao & Robson, 1974). Thus, item misfit is indicated by values of $\mathrm{S} - X^2_{RR}$ that are larger than the appropriate percentiles (say 95th or 99th percentile) of the χ^2_{J-1} distribution.

Simulation Studies

We performed a simulation study to evaluate the Type I error rates and power of the new $S-X^2_{RR}$ statistic defined in equation (24) and to compare its Type I error rates and power to those of the S-X^2 statistic (Orlando & Thissen, 2000) defined in equation (3). In the first part of the study, we compute and compare the Type I error rates of S-X^2 and $S-X^2_{RR}$ for data simulated from the 2PL model. In the second part of the simulation study, we examine and compare the power of S-X^2 and $S-X^2_{RR}$ for data simulated from the Rasch, the 2PL and the 3PL models. Both the statistics were computed using $\widehat{\boldsymbol{\eta}}$, which is the vector of the MMLEs of the item parameters.

Simulation Design

In the simulations, item scores were simulated under the Rasch, 2PL, and 3PL models. The test length was set as equal to 10, 20, or 40. The sample size was set equal to 500, 1000, 2000, or 4000. The true slope parameters, difficulty parameters, and guessing parameters were randomly generated from uniform distributions $U(1, 2)$, $U(-3, 3)$, and $U(0.05, 0.3)$, respectively, where, for example, $U(1, 2)$ denotes the uniform distribution between 1 and 2. Simulating the true parameter values from other distributions did not affect the comparative performance of the item-fit statistics. To investigate the Type I error rates of the two statistics, the data-generating model (the IRT model that was used to simulate the data) was fitted to the data. To investigate the power of the two statistics, the Rasch and 2PL models were fitted to data simulated from the 3PL model and the Rasch model was fitted to data simulated from the 2PL model. After the models were fitted to the data and the item fit statistics were computed, the Type I error rate of an item-fit statistic at the 5% significance level was computed as the proportion of values of the statistic that were larger than the 95th percentile of the χ^2 distribution with $J-1$ (for $S-X^2_{RR}$) or $J-L-1$ (for S-X^2) df for the simulation cases where the data-generating model and the fitted model were the same; the power of a statistic was computed as the proportion of values of the statistic that were larger than the 95th percentile of the χ^2 distribution with $J-1$ or $J-L-1$ df for the simulation cases where the data-generating model and the fitted model were different. Both Type I error rate and power for each combination of test length and sample size were computed from 100 replications. The true item parameters were resampled in each replication.

Results

Table 1 shows that the Type I error rates of the two statistics for the various simulation cases where the data-generating model and the fitted model were the same. The table shows that the Type I error rates of S-X^2 are larger than the nominal level in all simulation cases, a finding that is in agreement with findings on Type I error rates of S-X^2 in Glas and Suarez-Falcón (2003), Sinharay (2006), and Sinharay and Lu (2008). However, the Type I error rates of S-X^2 are not much larger than the nominal level for 40-item tests. The Type I error rates of the modified statistic $S-X^2_{RR}$ are considerably smaller than those of S-X^2 in all cases. Thus, $S-X^2_{RR}$ overcomes the Chernoff–Lehmann problem to a certain extent. However, the Type I error rates of $S-X^2_{RR}$ is considerably smaller than the nominal level for 10-item tests—we plan to investigate this issue in future research.

Table 2 shows the values of power of the two item-fit statistics for the various simulation cases where the data-generating model and the fitted model were different. The two columns with heading, for example, "2PL/1PL," show the power for the cases when the data were simulated from the 2PL model and analyzed using the Rasch model. Table 2 shows that the power of the modified statistic $S-X^2_{RR}$ is smaller than that of S-X^2. However, the slightly better

power of S-X^2 relative to $S-X^2_{RR}$ is likely a consequence of the inflated Type I error rate of the former statistic. As the sample size increases, the power of both statistics approach 1.0 for the "2PL/1PL" and "3PL/1PL" cases. The small power of both item statistics for the "3PL/2PL" case is an outcome of the fact that the 2PL model can explain data simulated from the 3PL model except for the case that the difficulty and guessing parameters for the latter model are too high (Sinharay, 2006).

Real Data Example

The two item-fit statistics, S-X^2 and $S-X^2_{RR}$, were computed for a real data set. The data set includes the item scores of 2000 examinees on a state test with 46 dichotomous and multiple-choice items (with 5 answer options for each item) designed to measure students' achievement in mathematics and was previously analyzed in Sinharay (2017).

The Rasch, 2PL, and 3PL models were fitted to the data set and the values of $S-X^2_{RR}$ and S-X^2 were computed for all items for each IRT model. Table 3 shows the number of items for which the item-fit statistics were statistically significant at the 5% level of significance for the three IRT models. The table shows that for each IRT model, the use of $S-X^2_{RR}$ leads to fewer items being identified as misfitting compared to that of S-X^2, with the difference being more prominent for the 2PL model. This finding agrees with the finding of smaller Type I error rate and power of $S-X^2_{RR}$ compared to S-X^2 in the simulation study. Although both statistics are significant for a considerable number of items for the Rasch and 2PL model, they are significant for only 6 and 3 items, respectively, for the 3PL model. Although the 3PL model seems to adequately fit the data set, more tests including tests for local independence (e.g., Chen & Thissen, 1997) and further investigations, should be conducted to finalize this conclusion.

The three panels of Figure 1 show scatter plots of S-X^2 versus $S-X^2_{RR}$ for the real data set under the three IRT models. The range of the X-axis is the same as that of the Y-axis in each panel. The range is much wider in the leftmost panel than in the other two panels. The panels include a diagonal line and also vertical and horizontal dashed lines indicating the critical values at 5% level of significance for the respective statistics. The last two panels show that for several items, S-X^2 is

Table 2. The Power of $S-X^2_{RR}$ and S-X^2 for Various Combinations of Data-generating Model and Fitted Model.

		2PL/1PL		3PL/1PL		3PL/2PL	
Test Length	Sample Size	S-X^2	$S-X^2_{RR}$	S-X^2	$S-X^2_{RR}$	S-X^2	$S-X^2_{RR}$
10	500	0.26	0.19	0.34	0.26	0.06	0.05
	1000	0.48	0.40	0.49	0.37	0.07	0.06
	2000	0.65	0.57	0.67	0.55	0.08	0.06
	4000	0.80	0.69	0.82	0.68	0.11	0.05
20	500	0.17	0.17	0.30	0.27	0.07	0.04
	1000	0.39	0.35	0.47	0.40	0.09	0.05
	2000	0.64	0.61	0.67	0.59	0.10	0.08
	4000	0.80	0.75	0.82	0.76	0.13	0.10
40	500	0.18	0.17	0.27	0.25	0.08	0.05
	1000	0.25	0.24	0.42	0.38	0.11	0.08
	2000	0.54	0.48	0.66	0.64	0.11	0.09
	4000	0.77	0.66	0.79	0.75	0.15	0.13

Table 3. The Number of Items with Statistically Significant Values of S-X^2 and S $-$ X^2_{RR} for the Three IRT models for the real data set.

Statistic	Rasch	2PL	3PL
S-X^2	33	18	6
S $-$ X^2_{RR}	31	12	3

Note. IRT = item response theory.

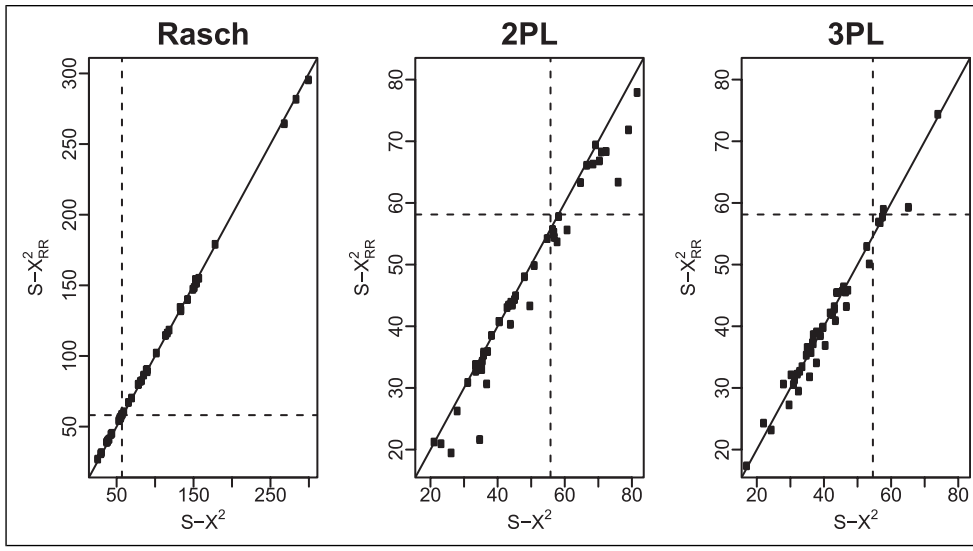

Figure 1. Plot of S-X^2 versus S $-$ X^2_{RR} for three item response theory models for the real data.

larger than its critical value, but S $-$ X^2_{RR} is smaller than its critical value. Because item misfit often leads to an item being removed from the item pool (Sinharay & Haberman, 2014) and items are costly, these results indicate that the use of S $-$ X^2_{RR} rather than S-X^2 may lead to considerable saving of resources in operational testing.

Conclusions and Recommendations

The item-fit statistic S-X^2 (Orlando & Thissen, 2000), in spite of its simplicity and popularity, does not have a known asymptotic null distribution (Sinharay, 2006) and the Type I error rate of the statistic is larger than the nominal level, especially for shorter tests. The present study adopts the modification procedure suggested by Rao and Robson (1974) to provide a modified version of S-X^2 that has a known χ^2 asymptotic null distribution. The statistic S-X^2 can be written as $\hat{\boldsymbol{\nu}}^T\hat{\boldsymbol{\nu}}$. The central idea of the modification of Rao and Robson (1974) is the computation of $\hat{\boldsymbol{\nu}}^T\Sigma_{\hat{\boldsymbol{\nu}}}^{-1}\hat{\boldsymbol{\nu}}$, where $\Sigma_{\hat{\boldsymbol{\nu}}}$ is an approximate variance-covariance matrix of $\hat{\boldsymbol{\nu}}$, so that $\hat{\boldsymbol{\nu}}^T\Sigma_{\hat{\boldsymbol{\nu}}}^{-1}\hat{\boldsymbol{\nu}}$ has a known χ^2 asymptotic null distribution. A major contribution of this paper is the derivation of the appropriate $\Sigma_{\hat{\boldsymbol{\nu}}}$. Thus, this paper suggests a χ^2-type statistic that (a) can be used to assess item fit for any IRT model for dichotomous items and (b) has a known asymptotic distribution under the null hypothesis. Item-fit statistics that have known asymptotic χ^2 distribution under the null hypothesis have been suggested for the Rasch model by, for example, Glas (1988), but there is a lack of such statistics for

non-Rasch IRT models. Thus, this paper makes an important contribution given that experts such as Box (1979) called for statistics that have known null distribution in assessing the fit of statistical models. Note that researchers such as Haberman et al. (2013) have suggested residual-based item-fit statistics that follow the standard normal distributions for non-Rasch IRT models, but we do not consider such statistics.

Simulation studies were conducted to compare the performance of S-X^2 and $S-X^2_{RR}$ with respect to Type I error rate and power. Results obtained from the simulation studies suggest that the Type I error rate of $S-X^2_{RR}$ is closer to the nominal level than S-X^2 across different conditions. However, $S-X^2_{RR}$ was found to be slightly conservative in comparison to S-X^2. Application of the two item-fit statistics to a real data set revealed that the number of misfitting items using $S-X^2_{RR}$ was smaller than that for S-X^2. In practice, item fit statistics such as $S-X^2_{RR}$ should be used along with other methods such as informative graphics and pair-wise item fit indexes in order to gain a thorough understanding of the type of misfit.

This paper has several limitations. First, it is possible to compare the two statistics for more simulated data and more real data. Second, the proposed statistic $S-X^2_{RR}$ applies only to dichotomous IRT models—it is possible to extend the statistic to tests with polytomous items or a mix of dichotomous and polytomous items in future research. Third, the current manuscript only investigates three unidimensional IRT models assuming the latent variable follows a normal distribution. To obtain better understanding of the suggested statistic, one can look into its performance in other cases including for non-normal ability distributions, multidimensional latent variables, and discrete latent variables. Finally, this manuscript only considers statistical significance and does not discuss practical significance on IRT model misfit (Hambleton & Han, 2005; Sinharay & Haberman, 2014).

Appendix A

Proof that $\mathrm{Cov}(O_{k_1}, O_{k_2})$ Computed at $\eta = \eta_0$ is Approximately Equal to Zero for Large Samples

Let S_i denote the total score of examinee i, who is randomly chosen from the hypothetical population of all possible examinees. Let us define an indicator variable W_{ik} as

$$W_{ik} = \begin{cases} 1, & S_i = k \\ 0, & S_i \neq k \end{cases}$$

Then O_k for an item of interest can be expressed as

$$O_k = \frac{\sum_i W_{ik} X_i}{\sum_i W_{ik}}$$

where X_i is the score of examinee i on the item.

Let us consider two possible values k_1 and k_2 of S_i, where $k_1 \neq k_2$, and define a vector $\boldsymbol{U}$ as

$$\boldsymbol{U} = \left(\sum_i W_{ik_1} X_i, \sum_i W_{ik_1}, \sum_i W_{ik_2} X_i, \sum_i W_{ik_2} \right)^{\top}$$

Then one can express O_{k_1} and O_{k_2} as

$$O_{k_1} = \frac{U_1}{U_2}, \quad O_{k_2} = \frac{U_3}{U_4}$$

where, for example, U_1 is the first component of $\boldsymbol{U}$. The Jacobian for the transformation from $\boldsymbol{O} = (O_{k_1}, O_{k_2})^\top$ to $\boldsymbol{U}$ is given by a matrix of the first derivatives of the elements of $\boldsymbol{O}$ with respect to those of $\boldsymbol{U}$, or, by

$$\boldsymbol{J} = \begin{bmatrix} J_1 & J_2 & 0 & 0 \\ 0 & 0 & J_3 & J_4 \end{bmatrix} \tag{A1}$$

where

$$J_1 = \frac{1}{U_2}, J_2 = -\frac{U_1}{U_2^2}, J_3 = \frac{1}{U_4}, J_4 = -\frac{U_3}{U_4^2} \tag{A2}$$

Consequently, using the multivariate delta method (e.g., Lehmann & Casella, 1998, p. 61), the variance-covariance matrix of $\boldsymbol{O}$ for large samples can be approximated as

$$\mathrm{Cov}(\boldsymbol{O}) \approx \tilde{\boldsymbol{J}} \boldsymbol{\Sigma}_{\boldsymbol{U}} \tilde{\boldsymbol{J}}^\top$$

where $\boldsymbol{\Sigma}_U$ is the variance-covariance matrix of the vector $\boldsymbol{U}$, $\tilde{\boldsymbol{J}}$ is the value of $\boldsymbol{J}$ provided in equation (A1) upon replacing the U_k's with their expected values computed at $\boldsymbol{\eta} = \boldsymbol{\eta_0}$, and the parameters $\boldsymbol{\eta}$ are fixed at $\boldsymbol{\eta}_0$. Using the result that the (i, j)-th element of the product of three matrices A, B and C is equal to the (matrix) product of the i-th row of A, the matrix B, and the j-th column of C (e.g., Banerjee & Roy, 2014, p. 12), the covariance between O_{k_1} and O_{k_2} can be approximated, for large samples, as

$$\mathrm{Cov}(O_{k_1}, O_{k_2}) \approx \left(\tilde{J}_1, \tilde{J}_2, 0, 0\right) \boldsymbol{\Sigma}_{\boldsymbol{U}} \left(0, 0, \tilde{J}_3, \tilde{J}_4\right)^\top$$

where $\tilde{J}_i$ is the value of J_i upon replacing the U_k's with their expected values computed at $\boldsymbol{\eta} = \boldsymbol{\eta_0}$, or, as

$$\mathrm{Cov}(O_{k_1}, O_{k_2}) \approx \tilde{J}_1 \tilde{J}_3 \sigma_{13} + \tilde{J}_1 \tilde{J}_4 \sigma_{14} + \tilde{J}_2 \tilde{J}_3 \sigma_{23} + \tilde{J}_2 \tilde{J}_4 \sigma_{24} \tag{A3}$$

where σ_{ij} is the (i, j)-th element of $\boldsymbol{\Sigma}_{\boldsymbol{U}}$.

One can compute σ_{24} as

$$\sigma_{24} = \mathrm{Cov}(U_2, U_4) = \mathrm{Cov}\left(\sum_i W_{ik_1}, \sum_i W_{ik_2}\right) = \sum_i \mathrm{Cov}(W_{ik_1}, W_{ik_2})$$

where the last equality holds because the item scores are independent over two different examinees i_1 and i_2, which results in $\mathrm{Cov}(W_{i_1 k_1}, W_{i_2 k_2}) = 0$. Consequently

$$\sigma_{24} = \sum_i [E(W_{ik_1} W_{ik_2}) - E(W_{ik_1}) E(W_{ik_2})] = -\sum_i E(W_{ik_1}) E(W_{ik_2}) \tag{A4}$$

because the raw score of examinee i cannot be equal to k_1 and also equal to k_2 so that $W_{ik_1} W_{ik_2}$ is equal to 0.

Now note that $E(W_{ik_1})$ is the probability that the raw score on the test is k_1 for an examinee who is randomly chosen from the population of all examinees, is equal to $\int S(T = k_1 | \theta, \boldsymbol{\eta}) \psi(\theta) \, d\theta$, and hence is the same over all the examinees. Therefore

$$E(W_{ik_1}) = \frac{1}{N}\sum_i E(W_{ik_1}) = \frac{1}{N}E\left(\sum_i W_{ik_1}\right) = \frac{1}{N}E(U_2) \tag{A5}$$

Similarly, one obtains

$$E(W_{ik_2}) = \frac{1}{N}E(U_4). \tag{A6}$$

Equations (A4) to (A6) imply that

$$\sigma_{24} = -\sum_i \left[\frac{1}{N}E(U_2)\right]\left[\frac{1}{N}E(U_4)\right] = -\frac{1}{N}E(U_2)E(U_4)$$

Let $\tilde{U}_k$ denote $E(U_k)$, where the expectation is computed at $\boldsymbol{\eta} = \boldsymbol{\eta_0}$, $k = 1, \ldots, 4$. Then

$$\sigma_{24} = -\frac{1}{N}\tilde{U}_2\tilde{U}_4 \tag{A7}$$

It is possible to prove in a similar manner that

$$\sigma_{13} = -\frac{1}{N}\tilde{U}_1\tilde{U}_3, \quad \sigma_{14} = -\frac{1}{N}\tilde{U}_1\tilde{U}_4, \quad \sigma_{23} = -\frac{1}{N}\tilde{U}_2\tilde{U}_3 \tag{A8}$$

Finally, equations (A2), (A3), (A7), and (A8) imply that

$$\begin{aligned}\mathrm{Cov}(O_{k_1}, O_{k_2}) \approx & -\tilde{J}_1\tilde{J}_3\frac{1}{N}\tilde{U}_1\tilde{U}_3 - \tilde{J}_1\tilde{J}_4\frac{1}{N}\tilde{U}_1\tilde{U}_4 - \tilde{J}_2\tilde{J}_3\frac{1}{N}\tilde{U}_2\tilde{U}_3 - \tilde{J}_2\tilde{J}_4\frac{1}{N}\tilde{U}_2\tilde{U}_4 \\ \approx & -\frac{1}{N}\left[\tilde{U}_1\tilde{U}_3\frac{1}{\tilde{U}_2}\frac{1}{\tilde{U}_4} - \tilde{U}_1\tilde{U}_4\frac{1}{\tilde{U}_2}\frac{\tilde{U}_3}{\tilde{U}_4^2} - \tilde{U}_2\tilde{U}_3\frac{\tilde{U}_1}{\tilde{U}_2^2}\frac{1}{\tilde{U}_4} + \tilde{U}_2\tilde{U}_4\frac{\tilde{U}_1}{\tilde{U}_2^2}\frac{\tilde{U}_3}{\tilde{U}_4^2}\right] = 0\end{aligned}$$

Acknowledgments

The authors would like to thank John Donoghue, Sooyeon Kim, Hongwen Guo, Lora Monfils, and two anonymous reviewers for several helpful comments that led to a significant improvement of the 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.

Author Note

Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

ORCID iDs

Sandip Sinharay https://orcid.org/0000-0003-4491-8510
Matthew S. Johnson https://orcid.org/0000-0003-3157-4165

Note

1. For example, the R package mirt (Chalmers, 2012) can be used to compute such a matrix.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Banerjee, S., & Roy, A. (2014). *Linear algebra and matrix analysis for statistics*. Chapman and Hall/CRC.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, *74*(365), 1–4. https://doi.org/10.1080/01621459.1979.10481600

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Scientific Software International.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Chernoff, H., & Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics*, *25*(3), 579–586. https://doi.org/10.1214/aoms/1177728726

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society*, *87*(3), 442–450.

Glas, C. A., & Suarez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. https://doi.org/10.1177/0146621602250530

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*(4), 525–546. https://doi.org/10.1007/bf02294405

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*(3), 417–440. https://doi.org/10.1007/s11336-012-9305-1

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking, & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Degnon Associates.

Harris, R. R., & Kanji, G. K. (1983). On the use of minimum chi-square estimation. *The Statistician*, *32*(4), 379. https://doi.org/10.2307/2987540

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*(3), 243–271. https://doi.org/10.2307/1164654

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kang, T., & Chen, T. T. (2010). Performance of the generalized S-X^2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89–96. https://doi.org/10.1007/s12564-010-9082-4

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer-Verlag.

Lim, H. (2020). irtplay: Unidimensional item response theory modeling. (R package version 1.6.2).

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*(4), 453–461. https://doi.org/10.1177/014662168400800409

Mislevy, R. J., & Bock, R. D. (1991). *BILOG 3.11 [computer software]*. Scientific Software International.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data [computer program]*. Scientific Software.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, *24*(1), 46–72. https://doi.org/10.1111/j.2517-6161.1962.tb00436.x

Rao, K. C., & Robson, D. S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics*, *3*(12), 1139–1153. https://doi.org/10.1080/03610917408548327

Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement*, *32*(5), 407–423. https://doi.org/10.1177/0146621607301278

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *The British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449. https://doi.org/10.1348/000711005x66888

Sinharay, S. (2017). How to compare parametric and nonparametric person-fit statistics using real data. *Journal of Educational Measurement*, *54*(4), 420–439. https://doi.org/10.1111/jedm.12155

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35. https://doi.org/10.1111/emip.12024

Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement*, *45*, 1–15. https://doi.org/10.1111/j.1745-3984.2007.00049.x

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631. https://doi.org/10.1177/0146621617707510

Stone, C. A. (2000). Monte Carlo based null distribtion for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*(1), 58–75. https://doi.org/10.1111/j.1745-3984.2000.tb01076.x

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory [computer software]*. Scientific Software International.

Zhang, B., & Stone, C. A. (2007). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, *68*(2), 181–196. https://doi.org/10.1177/0013164407301547