*ets research institute

NOVEMBER 2025
RR–25-13

RESEARCH REPORT

An Evaluation of Item Fit Based on Generalized Residual Item Response Functions

AUTHORS

Xiangyi Liao, Peter van Rijn, and Sandip Sinharay

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

An Evaluation of Item Fit Based on Generalized Residual Item Response Functions

Xiangyi Liao,[1] Peter van Rijn,[2] and Sandip Sinharay[3]
[1]University of British Columbia
[2]ETS Global, Amsterdam, The Netherlands
[3]ETS Research Institute, ETS, Princeton, New Jersey, United States

Abstract

Evaluation of item fit for item response theory (IRT) models often involves a comparison of the observed and expected item response functions (IRFs). Several statistics have been suggested for evaluating item fit based on the discrepancy between IRFs, but the asymptotic distributions of the statistics under the null hypothesis are often not well established. Haberman et al. developed a method for evaluating the fit of IRFs based on generalized residuals. These residuals are functions of the latent proficiency variable in the IRT model and follow the standard normal distribution asymptotically. We develop a method to summarize these generalized residuals into a single summary statistic for each item and evaluate its asymptotic distribution. Kondratek suggested a similar Wald-type statistic, but without accounting for the uncertainty in the estimation of the item parameters. Our method combines the work of Haberman and Kondratek, resulting in a single fit statistic per item while accounting for estimation error. A series of simulations was carried out to investigate the performance of our statistic and compare it to several popular item fit statistics. Our method resulted in similar Type I errors as Kondratek's statistic, with slightly better results in the case of small samples. Furthermore, the recovery was consistent across different levels of item difficulty, and power of the new item fit statistic was relatively low, except for problematic individual items, but this result was found with two competing statistics as well.

Keywords: Item fit; generalized residuals

Corresponding author: X. Liao, E-mail: xy.liao@ubc.ca

Introduction

In item response theory (IRT) modeling, item fit analysis is an important aspect in evaluating the accuracy of the item response function (IRF). Incorrect IRF specification can lead to incorrect scoring and fairness issues. An often-used method to evaluate item fit is to compare observed and expected IRFs. Several statistics have been developed to summarize the differences between IRFs, but null distributions are not available for some common cases, such as the INFIT (Wright & Panchapakesan, 1969) and root-mean-squared deviation (Oliveri & von Davier, 2011). Although methods like the bootstrap (Silva Diaz et al., 2022) and jackknife (Robitzsch, 2022) can overcome issues with asymptotics at the cost of computational time, it would be better to work with item fit statistics that have known asymptotic properties under a wide range of conditions.

In general, two main approaches can be distinguished in developing a chi-square statistic based on residuals between the observed and expected IRFs. Though both use a general framework whereby residuals are first computed by grouping individuals according to specific ranges of ability ("bins") and then summarized into a single χ^2 statistic, the definition of *bins* varies. In the first approach, IRFs are evaluated based on the estimate of the latent variable ($\hat{\theta}$) that explains the dependencies between items. Such indices include Bock's (1972) χ^2 statistic and Yen's (1981) Q_1 statistic. Their use of ability estimates in creating bins renders uncertainty in the true null distribution of the statistics, an ensuing problem with model-dependent statistics. A related consequence appears in their inflated Type I error and low power, particularly with short tests (Chon et al., 2010). The second approach, for example that used to derive Orlando and Thissen's (2000) S-X^2, avoids the use of $\hat{\theta}$ by grouping examinees based on their total score (i.e., the sum of scored responses to all items). While its Type I error rate is typically close to the nominal level, a concern with using S-X^2 is its low power (Stone & Zhang, 2003).

Haberman et al. (2013) developed a method for evaluating the fit of IRFs based on generalized residuals. This method produces asymptotically standard normal residuals as a function of the latent ability variable in the IRT model. The main goal of our research is to correctly combine these generalized residuals across the ability scale into a single summary statistic and establish its asymptotic distribution. Instead of using ability points to evaluate the IRFs, we make use of ability intervals, as Stone (2000) suggested. This results in a Wald-type test, which is assumed to have a chi-squared distribution. Recently, Kondratek (2022) conducted an extensive simulation study to evaluate a clever version of this statistic. However, his version

does not account for the fact that item parameters are estimated. We extend his statistic by accounting for the uncertainty in the estimation of item parameters.

A series of simulations is carried out to investigate the performance of this statistic under a variety of circumstances for the unidimensional two-parameter logistic (2PL) model (Birnbaum, 1968) and make comparisons to other commonly used item fit statistics. A benefit of using the theory behind generalized residuals is that, in principle, it can be easily extended, not only to other IRT models, including multigroup and multidimensional models, but also to models with response times (Sinharay & van Rijn, 2020).

Method

Item Response Theory Models

For a unidimensional 2PL model, the IRF for item j is given by

$$p(X_j = 1|\theta) = p_j(\theta) = \frac{\exp(a_j\theta + b_j)}{1 + \exp(a_j\theta + b_j)}, \tag{1}$$

where θ is the ability parameter, a_j is an item slope parameter, and b_j is an item intercept parameter. We also consider more complex models as the data-generating model, such as the 3PL and 4PL. The IRF of the unidimensional 4PL (Barton & Lord, 1981) is given by

$$p(X_j = 1|\theta) = p_j(\theta) = c_j + (d_j - c_j)\frac{\exp(a_j\theta + b_j)}{1 + \exp(a_j\theta + b_j)}, \tag{2}$$

where c_j and d_j are lower and upper asymptotes, respectively, to accommodate guessing and slipping behaviors on the test.

If a normal density $f(\theta)$ is assumed for the latent variable, the posterior of θ is

$$g(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)f(\theta)}{\int p(\mathbf{x}|\theta)f(\theta)d\theta}, \tag{3}$$

where $p(\mathbf{x}|\theta)$ is the likelihood of item response vector $\mathbf{x}$, typically under the assumption of local independence. In marginal maximum likelihood estimation of item parameters, the mean and standard deviation of $f(\theta)$ are typically fixed to 0 and 1, respectively, for the purpose of model identification, although identification issues may persist for the 3PL and 4PL models.

Derivations of New Item-Fit Statistic With Known Item Parameters

Let's first consider the case in which all parameters of a 2PL IRT model are known. Then, the expected IRF is given by Equation 1.

An alternative estimate of the IRF, or (pseudo-)observed IRF, is

$$\tilde{p}_j(\theta) = \frac{\sum_{i=1}^{N} x_{ij} g(\theta|\mathbf{x}_i)}{\sum_{i=1}^{N} g(\theta|\mathbf{x}_i)}, \tag{4}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots)$ is the item response vector of test taker i, $i = 1, \ldots, N$. The residual of the expected and observed IRFs, referred to as the residual IRF, is

$$r_j(\theta) = \tilde{p}_j(\theta) - p_j(\theta). \tag{5}$$

The estimated asymptotic variance of the residual IRF is

$$s^2\left[\tilde{p}_j(\theta)\right] = \frac{\sum_{i=1}^{N} \left\{g(\theta|\mathbf{x}_i)\left[x_{ij} - \tilde{p}_j(\theta)\right]\right\}^2}{\left[\sum_{i=1}^{N} g(\theta|\mathbf{x}_i)\right]^2}. \tag{6}$$

When the IRT model fits the data, the generalized residual IRF defined as $z_j(\theta) = [r_j(\theta)]/\{s[\tilde{p}_j(\theta)]\}$ converges in distribution to a standard normal variable (Haberman et al., 2013).

It would be tempting simply to take the sum of squared generalized residuals $z_j(\theta)$ over a selected number of θ points (e.g., the quadrature points) as a summary statistic, but such a sum does not have a known asymptotic null distribution. Kondratek (2022) used intervals based on θ ("bins") instead of θ points to construct observed and expected proportions correct at given intervals. Let $\Delta_{j1}, \ldots, \Delta_{jK}$ denote K nonintersecting intervals that cover the real line for item j. The intervals can be, for example, based on quantiles of the density $f(\theta)$, so that they are fixed across items, or adjusted by item difficulty to maintain roughly equal expected proportions for each interval. In the latter case, the intervals would be item specific. Such adaptive intervals can be constructed as in Kondratek (2022), where $n_{jk}E_{jk}(1 - E_{jk})$ is kept constant over k, with n_{jk} denoting the expected number of observations in the kth bin under $f(\theta)$ for item j and E_{jk} denoting the expected proportion of correct responses in the kth interval given by, respectively,

$$n_{jk} = N \int_{\Delta_{jk}} f(\theta) d\theta \tag{7}$$

$$E_{jk} = \frac{\int_{\Delta_{jk}} p_j(\theta) f(\theta) d\theta}{\int_{\Delta_{jk}} f(\theta) d\theta}. \tag{8}$$

We can then denote the conditional probability that θ falls into Δ_{jk} given response vector $\mathbf{x}_i$ as

$$\tau_{ijk} = \int_{\Delta_{jk}} g(\theta|\mathbf{x}_i)d\theta. \tag{9}$$

The observed proportion of correct responses in interval Δ_{jk} is then given by

$$O_{jk} = \frac{\sum_{i=1}^{N} x_{ij}\tau_{ijk}}{\sum_{i=1}^{N} \tau_{ijk}}. \tag{10}$$

An alternative expected proportion of correct responses for item j can be based on the item response vector without item j, denoted by $\mathbf{x}_{i\setminus j}$ (see, e.g., Kondratek, 2022, Equation 12). It can be argued that doing so purifies the calculation of the expected proportion of the scrutinized item, which results in

$$e_{ijk} = \int_{\Delta_{jk}} p_j(\theta)g(\theta|\mathbf{x}_{i\setminus j})d\theta \tag{11}$$

$$E^*_{jk} = \frac{\sum_{i=1}^{N} e_{ijk}\tau_{ijk}}{\sum_{i=1}^{N} \tau_{ijk}}. \tag{12}$$

In calculating τ_{ijk}, one would have to then also use $g(\theta|\mathbf{x}_{i\setminus j})$ instead of $g(\theta|\mathbf{x}_i)$. However, Kondratek does not mention this point, and because excluding the item complicates the computation, we retain it for simplicity.

Asymptotic normality of the O_{jk} holds if the observed and expected frequencies (i.e., the numerators in Equations 10 and 12) are not too small. One typical rule of thumb is that both $n_{jk}E_{jk}$ and $n_{jk}(1 - E_{jk})$ should be larger than 20 (Kondratek, 2022). The covariance matrix $\mathbf{V}_j$ of $\mathbf{O}_j$ has elements

$$v_{jkl} = \frac{\sum_{i=1}^{N} \tau_{ik}\tau_{il}(x_{ij} - O_{jk})(x_{ij} - O_{jl})}{\sum_{i=1}^{N} \tau_{ik} \sum_{i=1}^{N} \tau_{il}}. \tag{13}$$

Kondratek then defined a Wald-like item fit statistic as

$$X^2_{W_j} = (\mathbf{O}_j - \mathbf{E}_j)' \mathbf{V}_j^{-1} (\mathbf{O}_j - \mathbf{E}_j), \tag{14}$$

which has an asymptotic chi-squared distribution with K degrees of freedom.

Derivations of New Item-Fit Statistic With Estimated Item Parameters

Under a 2PL model, the expected IRF with estimated item parameters is

$$\hat{p}_j(\theta) = \frac{\exp(\hat{a}_j\theta + \hat{b}_j)}{1 + \exp(\hat{a}_j\theta + \hat{b}_j)}. \tag{15}$$

Again, the alternative estimate of the IRF is

$$\bar{p}_j(\theta) = \frac{\sum_{i=1}^{N} x_{ij}\hat{g}(\theta|\mathbf{x}_i)}{\sum_{i=1}^{N} \hat{g}(\theta|\mathbf{x}_i)}, \tag{16}$$

where $\hat{g}(\theta|\mathbf{x}_i)$ is the estimated conditional density of θ given item response vector $\mathbf{x}_i$ using estimated item parameters. The residual IRF is

$$\hat{r}_j(\theta) = \bar{p}_j(\theta) - \hat{p}_j(\theta). \tag{17}$$

Haberman et al. (2013) noted that the estimate in Equation 16 is a ratio so that standard formulas can be applied to obtain the variance of the residual. This variance is found in Equation 46 of their paper:

$$s^2(\hat{r}_j(\theta)) = \frac{\sum_{i=1}^{N}\left[\hat{g}(\theta|\mathbf{x}_i)\left[x_{ij} - \hat{p}_j(\theta)\right] - \left[\hat{\mathbf{c}}_j(\theta)\right]'\nabla\ell_i(\hat{\boldsymbol{\xi}})\right]^2}{\left[\sum_{i=1}^{N}\hat{g}(\theta|\mathbf{x}_i)\right]^2}, \tag{18}$$

where

$$\hat{\mathbf{c}}_j(\theta) = N^{-1}\bar{\mathbf{J}}^{-1}\sum_{i=1}^{N}\hat{g}(\theta|\mathbf{x}_i)\left[x_{ij} - \hat{p}_j(\theta)\right]\nabla\ell_i(\hat{\boldsymbol{\xi}}) \tag{19}$$

and

$$\bar{\mathbf{J}} = N^{-1}\sum_{i=1}^{N}\nabla\ell_i(\hat{\boldsymbol{\xi}})\left[\nabla\ell_i(\hat{\boldsymbol{\xi}})\right]'. \tag{20}$$

When the IRT model fits the data, the generalized residual IRF $\hat{z}_j(\theta) = [\hat{r}_j(\theta)]/\{s[\hat{r}_j(\theta)]\}$ converges in distribution to a standard normal variable (Haberman et al., 2013).

We aim to find the $K \times K$ covariance matrix of the difference between the observed and expected proportion of correct responses in the intervals for estimated item parameters denoted by $\hat{\mathbf{V}}_j$. We denote the conditional probability that θ falls into Δ_{jk} given response vector $\mathbf{x}_i$ using estimated item parameters as

$$\hat{\tau}_{ijk} = \int_{\Delta_{jk}} \hat{g}(\theta|\mathbf{x}_i)d\theta. \tag{21}$$

The observed proportion of correct responses in interval Δ_{jk} is then given by

$$\hat{O}_{jk} = \frac{\sum_{i=1}^{N} x_{ij}\hat{\tau}_{ijk}}{\sum_{i=1}^{N}\hat{\tau}_{ijk}}. \tag{22}$$

The expected proportion of correct responses in interval Δ_{jk} is

$$\hat{E}_{jk} = \int_{\Delta_{jk}} \hat{p}_j(\theta)f(\theta)d\theta, \tag{23}$$

where $f(\theta)$ is the standard normal density (i.e., the mean and standard deviation are fixed for model identification). Again, asymptotic normality of the $\hat{O}_{jk}$ holds if the observed and expected frequencies are not too small. The estimated asymptotic covariance matrix $\hat{\mathbf{V}}_j$ of $\hat{\mathbf{O}}_j - \hat{\mathbf{E}}_j$ has elements (k, l):

$$\hat{v}_{jkl} = \frac{\sum_{i=1}^{N} \hat{\tau}_{ijk}\hat{\tau}_{ijl}\left[x_{ij} - \hat{O}_{jk} - \hat{\mathbf{c}}'_{jk}\nabla\ell_i(\hat{\boldsymbol{\xi}})\right]\left[x_{ij} - \hat{O}_{jl} - \hat{\mathbf{c}}'_{jl}\nabla\ell_i(\hat{\boldsymbol{\xi}})\right]}{\sum_{i=1}^{N}\hat{\tau}_{ijk}\sum_{i=1}^{N}\hat{\tau}_{ijl}}, \tag{24}$$

where

$$\hat{\mathbf{c}}'_{jk} = N^{-1}\bar{\mathbf{J}}^{-1}\sum_{i=1}^{N}\left[x_{ij} - \hat{O}_{jk}\right]\nabla\ell_i(\hat{\boldsymbol{\xi}}) \tag{25}$$

and $\hat{\mathbf{c}}'_{jl}$ is defined analogously.

The item fit statistic for estimated item parameters is then defined by

$$\hat{X}^2_{W_j} = \left(\hat{\mathbf{O}}_j - \hat{\mathbf{E}}_j\right)' \hat{\mathbf{V}}_j^{-1}\left(\hat{\mathbf{O}}_j - \hat{\mathbf{E}}_j\right), \tag{26}$$

which has an asymptotic chi-squared distribution with $K - p$ degrees of freedom, where p is the number of estimated item parameters (i.e., 2 in the case of the 2PL).

Simulation

We conducted two simulation studies to evaluate the Type I error and power of our statistic $\hat{X}^2_W$ in comparison with X^2_W (note that we drop the item index j for convenience). Both statistics are computed by plugging in the item parameter estimates as if they were the true parameters, but X^2_W does not correct for the uncertainty in parameter estimates and is close to Kondratek's (2022) statistic. In addition, two other well-known item-fit statistics were used in the comparison—Orlando and Thissen's (2000) S-X^2 and Yen's (1981) Q_1—and we used the R package MIRT (Chalmers, 2012) to calculate them.

Simulation 1: Type I Error

In the first simulation study, we manipulated three factors: (a) number of θ intervals (three and five adaptive bins), (b) test length (20 and 40 items), and (c) sample size (200, 500, and 1,000). The smaller sample of 200 was intended to explore how the fit statistics were affected by an increased amount of uncertainty in item parameter estimates. The three factors were crossed with each other.

Table 1. Type I Error (%) of $\hat{X}^2_W$, X^2_W, S-X^2, and Q_1, 200 Replications

K	Test length (items)	Sample size	$\hat{X}^2_W$	X^2_W	S-X^2	Q_1
3	20	200	5.60	5.50	4.56	6.80
		500	5.90	5.90	4.90	18.50
		1,000	5.40	5.35	5.00	52.17
	40	200	6.10	5.97	4.27	4.43
		500	4.98	4.92	4.85	5.91
		1,000	5.40	5.40	4.90	8.49
5	20	200	7.70	7.45	4.56	6.80
		500	6.45	6.45	4.90	18.50
		1,000	5.00	5.00	5.00	52.17
	40	200	7.95	7.75	4.27	4.43
		500	5.82	5.78	4.85	5.91
		1,000	5.78	5.73	4.90	8.49

We generated item response data using the unidimensional 2PL model. The item parameters were generated from the distributions $\log a_j \sim \mathrm{N}(0, 0.25)$ and $b_j \sim \mathrm{N}(0, 0.8)$ and the respondent proficiencies were generated from a standard normal distribution. The response data were then calibrated with the 2PL model, and all four item fit statistics were calculated. For each of the 12 simulation conditions, 200 replications were performed. The empirical Type I error rate for each individual statistic was computed by averaging the proportions significant across all items and replications.

Table 1 shows the Type I error rates for the simulation conditions at the 5% level. In contrast to the high Type I error rate for Q_1 (unreasonably high, given shorter tests), our statistic $\hat{X}^2_W$ is comparable to S-X^2 in regards to the Type I error being close to the nominal level. Also, the Type I error rates of $\hat{X}^2_W$ are similar to those of X^2_W, indicating that accounting for the uncertainty in the estimated parameters did not inflate the error rate.

We further examined the relationship between item difficulty and the p-value of the $\hat{X}^2_W$ statistic for the simulated data to understand the impact of item difficulty and the use of adaptive bins on Type I error.

Figure 1 shows how the p-value of $\hat{X}^2_W$ is related to item difficulty using spline smoothing. Under the 2PL model, the p-values should have a uniform distribution so that the smoothed average is expected to be around .5. Although the adaptive bins aim to maintain constant

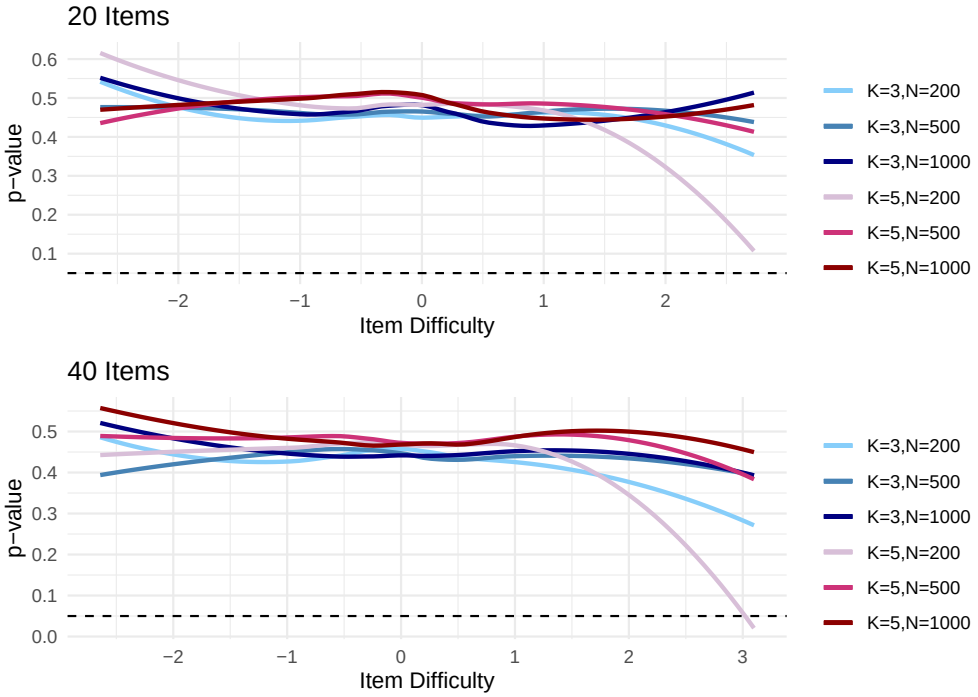

Figure 1. Relationship between item difficulty and p-value of $\hat{X}^2_W$. Solid lines represent the smoothed functions of p-value in relation to item difficulty across items in all 200 replications under different conditions, whereas the dashed line is an extreme p-value level .05.

proportions within each interval, Type I errors were largely inflated for $N = 200$, presumably because of a small number of respondents in some intervals for very difficult items.

Simulation 2: Power

The second simulation study compared the power of the fit statistics. Because the Type I error rate of Yen's (1981) Q_1 is excessively high (see Table 1), we excluded it from the power analysis.

We considered similar factors as in the previous analysis and additionally manipulated the generating model (GM) to detect misfit. Specifically, we used the following three scenarios to introduce misfit. In the first scenario, the calibrating model (CM) has fewer parameters than the GM, where we consider more complex 3PL and 4PL models as the GM. The IRF of the unidimensional 4PL model (Barton & Lorde, 1981) is given by

$$\hat{p}_j(\theta) = c_j + (d_j - c_j)\frac{\exp(\hat{a}_j\theta + \hat{b}_j)}{1 + \exp(\hat{a}_j\theta + \hat{b}_j)}, \tag{27}$$

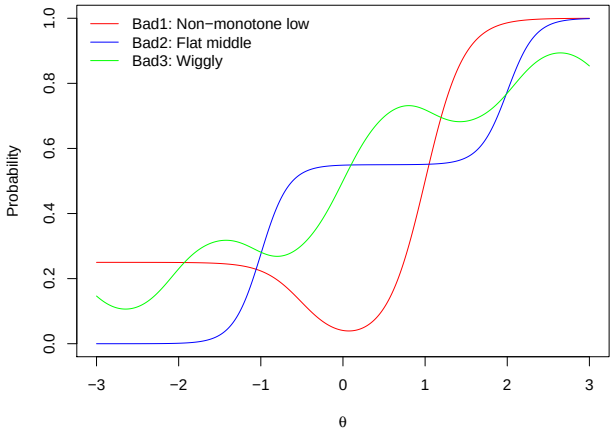

Figure 2. Item response functions for three types of misfitting items.

where c_j and d_j are lower and upper asymptotes, respectively, to accommodate guessing and slipping behaviors on the test. If $d_j = 1$, the preceding IRF reduces to that for the 3PL model. In this scenario, the GM can be either the 3PL model with $\log a_j \sim \mathrm{N}(0, 0.25)$, $b_j \sim \mathrm{N}(0, 0.8)$ and $c_j \sim \mathrm{Beta}(8, 32)$ or the 4PL model with $\log a_j \sim \mathrm{N}(0, 0.25)$, $b_j \sim \mathrm{N}(0, 0.8)$, $c_j \sim \mathrm{Beta}(8, 32)$, and $d_j \sim \mathrm{Beta}(32, 8)$. The second scenario has a proportion of only 10% of misfitting items, with the data for misfitting items being generated from either the 3PL or 4PL model. In the third scenario, we add three types of bad items one at a time, which is similar to the approach used by Sinharay (2006) and van Rijn et al. (2016). These bad items are as follows:

- BAD1: Nonmonotone IRF in the lower region of θ

$$P(Y = 1|\theta) \equiv \frac{1}{4}\mathrm{logit}^{-1}(-4.25(\theta + 0.5)) + \mathrm{logit}^{-1}(4.25(\theta - 1))$$

- BAD2: Flat IRF in the middle region of θ

$$P(Y = 1|\theta) \equiv 0.55\mathrm{logit}^{-1}(5.95(\theta + 1)) + 0.45\mathrm{logit}^{-1}(5.95(\theta - 2))$$

- BAD3: Wiggly, nonmonotone IRF

$$P(Y = 1|\theta) \equiv 0.65\mathrm{logit}^{-1}(1.5\theta) + 0.35\mathrm{logit}^{-1}(\sin(3\theta))$$

Figure 2 presents the corresponding IRFs for the three types of misfitting items. Regardless of the various GMs, we fit the 2PL model to the data.

Table 2. Power Rates (%) of $\hat{X}_W^2$, X_W^2, and $S\text{-}X^2$, 200 Replications

K	Test length (items)	Sample size	$\hat{X}_W^2$	X_W^2	$S\text{-}X^2$
		Simulated 4PL, estimated 2PL			
3	20	200	6.45	6.20	4.20
		500	6.50	6.40	6.00
		1,000	7.30	7.30	5.70
	40	200	6.48	6.45	5.22
		500	6.75	6.73	4.72
		1,000	9.32	9.30	5.73
5	20	200	7.10	6.95	4.20
		500	6.75	6.70	6.00
		1,000	6.30	6.20	5.70
	40	200	7.05	6.82	5.22
		500	6.48	6.40	4.72
		1,000	8.25	8.25	5.73
		Simulated 3PL, estimated 2PL			
3	20	200	11.45	11.30	4.70
		500	17.20	17.15	6.00
		1,000	29.20	29.15	6.85
	40	200	10.40	10.22	5.12
		500	15.72	15.65	5.40
		1,000	25.65	25.65	7.18
5	20	200	11.50	11.15	4.70
		500	15.90	15.75	6.00
		1,000	26.45	26.25	6.85
	40	200	11.33	11.15	5.12
		500	14.70	14.47	5.40
		1,000	24.52	24.47	7.18

Note. 2PL = two-parameter logistic. 3PL = three-parameter logistic. 4PL = four-parameter logistic.

Table 2 shows the power rates of the three item fit statistics if the GM is either the 4PL model or the 3PL model and the CM is the 2PL model. In general, the power is low for detecting this type of misfit. The $\hat{X}_W^2$ statistic provides slightly larger power rates for small samples than does X_W^2, and both produce larger rates than does $S\text{-}X^2$.

Table 3 shows the power and false alarm rates of the three item fit statistics if, for 10% of the items, the GM is either the 4PL model or the 3PL model and the CM is the 2PL model. The power is in general smaller when the GM is the 4PL model than when it is the 3PL model. It may seem paradoxical that the 2PL model fits data generated under the 4PL better than it does under the 3PL. However, the IRF is more symmetric for the 4PL model that we considered (i.e.,

Table 3. Power and False Alarm Rates (%) of $\hat{X}^2_W$, X^2_W, and S-X^2 for 10% Misfitting Items, 200 Replications

			Power			False Alarm Rate		
K	Test length (items)	Sample size	$\hat{X}^2_W$	X^2_W	S-X^2	$\hat{X}^2_W$	X^2_W	S-X^2
		Simulated 10% 4PL, estimated 2PL						
3	20	200	6.00	6.00	7.50	5.89	5.72	4.65
		500	9.00	9.00	7.00	5.72	5.67	5.44
		1,000	8.50	8.50	10.00	5.06	5.00	5.61
	40	200	7.00	7.00	4.00	6.17	6.03	4.22
		500	8.00	7.75	6.75	5.28	5.19	4.83
		1,000	12.00	11.75	8.25	5.50	5.50	5.56
5	20	200	6.50	6.50	7.50	6.89	6.78	4.65
		500	8.00	8.00	7.00	5.39	5.28	5.44
		1,000	8.50	8.50	10.00	5.56	5.56	5.61
	40	200	8.75	8.75	4.00	8.08	7.86	4.22
		500	11.75	11.75	6.75	5.33	5.28	4.83
		1,000	10.75	10.75	8.25	5.89	5.86	5.56
		Simulated 10% 3PL, estimated 2PL						
3	20	200	9.00	8.50	8.50	6.22	6.22	4.88
		500	14.50	14.50	9.50	6.00	5.83	5.67
		1,000	20.00	20.00	12.00	5.61	5.61	5.39
	40	200	8.00	8.00	6.50	5.89	5.72	4.30
		500	12.00	12.00	9.25	5.61	5.53	5.31
		1,000	22.50	22.50	8.25	5.78	5.72	5.31
5	20	200	10.00	9.50	8.50	7.17	6.94	4.88
		500	14.00	14.00	9.50	5.44	5.39	5.67
		1,000	16.00	16.00	12.00	5.61	5.61	5.39
	40	200	10.50	10.00	6.50	8.00	7.81	4.30
		500	15.50	15.00	9.25	5.78	5.67	5.31
		1,000	22.75	22.75	8.25	6.14	6.11	5.31

Note. 2PL = two-parameter logistic. 3PL = three-parameter logistic. 4PL = four-parameter logistic.

the average c parameter is roughly 0.20 and the average d parameter is roughly 0.80) than for the 3PL model (where the average c parameter is roughly 0.20 and the upper asymptote is fixed at 1). As a result, the symmetric 2PL model can adapt more easily to scores simulated from the 4PL model. In general though, the power is low in all the conditions considered in Table 3. Type I error rates rates are not that different from those found in the first simulation study (see Table 1).

In Table 4, the power and false alarm rates of the three statistics for the three types of misfitting items are shown. The power and false alarm rates of all three statistics are quite satisfactory for the first bad-item type, except perhaps the false alarm rates with five bins and 200

Table 4. Power and False Alarm Rates (%) of $\hat{X}^2_W$, X^2_W, and S-X^2 for Three Types of Misfitting Items, 200 Replications

			Power			False alarm rate		
K	Test length (items)	Sample size	$\hat{X}^2_W$	X^2_W	S-X^2	$\hat{X}^2_W$	X^2_W	S-X^2
		BAD1: Nonmonotone IRF in the lower region of θ						
3	20	200	96.00	96.00	79.00	5.53	5.26	4.24
		500	100.00	100.00	100.00	5.74	5.68	5.32
		1,000	100.00	100.00	100.00	5.68	5.58	5.95
	40	200	99.00	99.00	87.00	5.92	5.79	4.50
		500	100.00	100.00	100.00	5.13	5.08	4.55
		1,000	100.00	100.00	100.00	5.59	5.51	5.33
5	20	200	96.00	96.00	79.00	7.58	7.21	4.24
		500	100.00	100.00	100.00	6.16	6.11	5.32
		1,000	100.00	100.00	100.00	5.53	5.47	5.95
	40	200	100.00	100.00	87.00	7.77	7.56	4.50
		500	100.00	100.00	100.00	5.79	5.74	4.55
		1,000	100.00	100.00	100.00	5.67	5.59	5.33
		BAD2: Flat IRF in the middle region of θ						
3	20	200	33.00	33.00	16.00	5.21	5.11	5.17
		500	70.00	70.00	28.00	5.32	5.26	4.47
		1,000	96.00	96.00	63.00	5.47	5.47	5.63
	40	200	48.00	48.00	17.00	6.08	5.79	4.42
		500	83.00	83.00	44.00	5.03	4.95	5.11
		1,000	99.00	99.00	86.00	5.54	5.51	4.46
5	20	200	36.00	34.00	16.00	7.26	7.16	5.17
		500	76.00	76.00	28.00	5.16	5.11	4.47
		1,000	97.00	97.00	63.00	5.58	5.58	5.63
	40	200	57.00	56.00	17.00	7.33	7.10	4.42
		500	91.00	91.00	44.00	5.69	5.54	5.11
		1,000	99.00	99.00	86.00	5.56	5.54	4.46
		BAD3: Wiggly, nonmonotone IRF						
3	20	200	6.00	6.00	8.00	5.53	5.47	4.82
		500	7.00	7.00	16.00	5.26	5.21	4.11
		1,000	8.00	8.00	26.00	5.00	4.95	4.58
	40	200	4.00	4.00	7.00	6.21	6.05	3.75
		500	8.00	8.00	16.00	4.92	4.85	4.66
		1,000	9.00	9.00	33.00	5.67	5.67	4.62
5	20	200	9.00	9.00	8.00	7.42	7.32	4.82
		500	22.00	22.00	16.00	5.16	5.16	4.11
		1,000	40.00	40.00	26.00	5.21	5.21	4.58
	40	200	7.00	7.00	7.00	7.74	7.59	3.75
		500	34.00	34.00	16.00	5.46	5.36	4.66
		1,000	69.00	69.00	33.00	5.28	5.26	4.62

Note. IRF = item response function.

respondents. For the second bad-item type, the power is lower for smaller samples and for S-X^2. The power for $\hat{X}^2_W$ is slighter larger than that for X^2_W for five bins and 200 respondents. For the third bad-item type, the power is generally not great, except for $\hat{X}^2_W$ and X^2_W with five bins, 40 items, and 1,000 respondents.

Discussion

Our method combines ideas from the work of Haberman et al. (2013) and Kondratek (2022), resulting in a single fit statistic per item while accounting for estimation error of item parameters. The Type I error rates of our suggested statistic were similar to those of Kondratek's statistic, even with small samples, in which case, item parameter uncertainty is substantial. In addition, the power of our item fit statistic was slightly larger compared to Kondratek's statistic and considerably larger compared to Orlando and Thissen's (2000) S-X^2 statistic. However, the power to detect misfit was generally low when data were simulated from the 3PL and 4PL models and analyzed using the 2PL model. Also, wiggly IRFs were hard to detect using all item-fit statistics considered in this paper.

Although the results are generally promising, our simulations were limited in terms of the chosen IRT models and the comparisons that were made. So further work is needed in this area. In general, most item fit statistics can have problems in specific situations (e.g., when there are floor or ceiling effects and the IRF is relatively flat), and our method is no exception. Nevertheless, our work contributes to the search for item fit statistics with good asymptotic properties that do not require additional computational burden to find the null distribution. In the future, we plan to extend the method to other IRT models, such as multidimensional IRT models and multigroup IRT models.

References

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model* (Research Report No. RR-81-20). ETS. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. https://doi.org/10.1007/BF02291411

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

Chon, K. H., Lee, W.-C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement, 47*(3), 318–338. https://doi.org/10.1111/j.1745-3984.2010.00116.x

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

Kondratek, B. (2022). Item-fit statistic based on posterior probabilities of membership in ability groups. *Applied Psychological Measurement, 46*(6), 462–478. https://doi.org/10.1177/01466216221108061

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315.

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

Robitzsch, A. (2022). Statistical properties of estimators of the RMSD item fit statistic. *Foundations, 2*(2), 488–503. https://doi.org/10.3390/foundations2020032

Silva Diaz, J. A., Kohler, C., & Hartig, J. (2022). Performance of INFIT and OUTFIT confidence intervals calculated via parametric bootstrapping. *Applied Measurement in Education, 35*(2), 116–132. https://doi.org/10.1080/08957347.2022.2067540

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*(2), 429–449. https://doi.org/10.1348/000711005X66888

Sinharay, S., & van Rijn, P. W. (2020). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics, 45*(5), 534–568. https://doi.org/10.3102/1076998620911935

Stone, C. A. (2000). Monte Carlo based null distribution 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

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education, 4*, Article 10. https://doi.org/10.1186/s40536-016-0025-3

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23–48. https://doi.org/10.1177/001316446902900102

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Suggested Citation:

Liao, X., van Rijn, P., & Sinharay, S. (2025). *An evaluation of item fit based on generalized residual item response functions* (Research Report No. RR-25-13). ETS. https://doi.org/10.64634/b68vz316

Action Editor: Tim Davey

Reviewers: Paul Jewsbury and Yi-Hsuan Lee