Item-Fit Statistic Based on Posterior Probabilities of Membership in Ability Groups

Applied Psychological Measurement
2022, Vol. 46(6) 462–478
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216221108061
journals.sagepub.com/home/apm
SAGE

Bartosz Kondratek

Abstract

A novel approach to item-fit analysis based on an asymptotic test is proposed. The new test statistic, χ^2_w, compares pseudo-observed and expected item mean scores over a set of ability bins. The item mean scores are computed as weighted means with weights based on test-takers' *a posteriori* density of ability within the bin. This article explores the properties of χ^2_w in case of dichotomously scored items for unidimensional IRT models. Monte Carlo experiments were conducted to analyze the performance of χ^2_w. Type I error of χ^2_w was acceptably close to the nominal level and it had greater power than Orlando and Thissen's $S-\text{x}^2$. Under some conditions, power of χ^2_w also exceeded the one reported for the computationally more demanding Stone's χ^{2*}.

Keywords

item response theory, item response theory model fit, item-fit, asymptotic test

Introduction

Item response theory (IRT) models are a potent tool for explaining test behavior. However, the validity of analyses that involve IRT is critically related to the extent to which an IRT model fits the data. Several authors pointed to consequences of lack of fit of the IRT model for subsequent analyses (e.g., Wainer & Thissen, 1987; Woods, 2008; Bolt et al., 2014). *The Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014) recommend that provision of evidence of model fit should be a prerequisite for making any inferences based on IRT.

Analysis of fit at a level of single item plays an especially important role in the assessment of IRT model validity, since IRT models are designed for the very purpose of explaining observable data by separating item properties from the properties of the test-takers. In unidimensional IRT models for dichotomous items, the probability of the response pattern $\boldsymbol{y} = (y_1,\ldots, y_n)$ conditional on ability of the test-taker, θ, is assumed to follow

Institute for Educational Research, Warsaw, Poland

Corresponding Author:
Bartosz Kondratek, Institute for Educational Research, Górczewska 8, Marzanny 01-180, Poland.
Email: bartosz.kondratek@gmail.com

$$p(\boldsymbol{y}|\theta) = \prod_{j=1}^{n} (f_j(\theta))^{y_j} (1 - f_j(\theta))^{1-y_j}, \tag{1}$$

where f_j is a monotonically increasing function that describes the conditional probability of a correct response to item $j \in \{1,\ldots,n\}$. The marginal likelihood of response vector y is given by

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\theta) g(\theta) d\theta, \tag{2}$$

where g is the a priori ability distribution. Finally, *a posteriori* density of θ given response vector $\mathbf{y}$ is

$$g(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta) g(\theta)}{p(\boldsymbol{y})} \cdot \tag{3}$$

These equations illustrate how item response functions f_j mirror the structure of observable data. They serve as building blocks of the whole IRT model, and their form impacts any inferences regarding test-taker position on the θ continuum. Therefore, item-fit analysis is crucial in IRT. The item-level misfit information allows one to improve the overall model fit by discarding the misfitting items from analyses or by replacing the IRT model with one defined over a richer parameter space.

Many different approaches to item-fit testing have been developed. Despite ample research on the topic, the available solutions either are restricted to special cases of models and testing designs or require resampling. This article describes a universal and computationally feasible method for testing item-fit that aims at filling the gap in what is currently proposed.

Existing Item-Fit Statistics

Item-fit statistics are measures of discrepancy between the expected item performance, based on f_j, and the observed item performance. Usually, the difference between the observed and expected item-score is calculated over groups of test-takers with similar ability and aggregated into a single number. The pursuit of an item-fit measure that would allow for statistical testing started with the first applications of IRT. A selective account of previous research is presented to provide the context for the development of the approach proposed in the article. An in-depth review of research on item-fit is available, for example, in Swaminathan et al. (2007).

Grouping on Point Estimates of θ

First advancements in this field of research were inspired by the solutions available for models that did not deal with latent variables. These early approaches grouped test-takers on their point estimates of θ and computed Pearson's X^2 (Bock, 1972; Yen, 1981) or likelihood-ratio test statistic G^2 (McKinley & Mills, 1985). Uncertainty in measurement of θ is not accounted for under such grouping and the observed counts within groups are treated as independent from each other. Consequently, these fit statistics produce inflated Type I error rates, especially when tests are short (Orlando & Thissen, 2000; Stone & Hansen, 2000). Only in the case of Rasch-family of models, where the number-correct score is a sufficient statistic for $\widehat{\theta}$, such approaches yield item-fit statistics that are in accordance with the postulated asymptotic distribution (Andersen, 1973; Glas & Verhelst, 1989).

Grouping on Observed Sum-Scores

Orlando and Thissen (2000) approached the problem from a different angle. Instead of relying on the partition of latent trait, they grouped test-takers on their number-correct scores. This allowed to compute the observed frequency of correct responses directly from observable data. In order to compute the expected frequency of correct responses at a given score group, Orlando and Thissen ingeniously employed the algorithm of Lord and Wingersky (1984). Their version of Pearson's statistic, $\mathrm{S}-\mathrm{X}^2$, has become a standard point of reference in studies on item-fit because $\mathrm{S}-\mathrm{X}^2$ is fast to compute and has Type I error rates very close to the nominal level.

A likelihood-based approach to item-fit with aggregation over sum-scores was developed by Glas (1999) and further expanded by Glas and Suarez-Falcon (2003). This approach stands out from other item-fit measures, not only by accounting for the stochastic nature of item parameters, but also because it does not directly bank on the observed versus expected difference. Item misfit is modeled by additional group-specific parameters, *modification indices*, introduced in order to capture systematic deviance of data from the item response function. To test for significance, the modification indices are compared to zero and the Lagrange multiplier test is employed. However, as pointed by Sinharay (2006), results of simulations run by Glas and Suarez-Falcon (2003) show that Type I error of their statistic can be elevated under some score groupings.

Resampling Methods

Partitioning of ability on observed scores, rather than latent θ, limits practical applications of statistics such as $\mathrm{S}-\mathrm{X}^2$ or LM proposed by Glas. When test-takers respond to different sets of items their raw sum-scores become incomparable. However, a lot of appeal of IRT arises exactly from the fact that it can be applied in analysis of incomplete testing designs. So, the pursuit for a solution that relies on residuals computed over θ scale has never stopped.

Stone (2000) developed a simulation-based approach. He proposed a χ^{2*} statistic calculated over quadrature points of θ, with a resampling algorithm for determining the distribution of χ^{2*} under the null hypothesis. Stone's χ^{2*} repeatedly proved to provide acceptable Type I error rate and it exceeded $\mathrm{S}-\mathrm{X}^2$ in power (Stone & Zhang, 2003; Chon et al., 2010; Chalmers & Ng, 2017). However, this came at a significant computational cost.

Other computationally intensive approaches were also developed. Sinharay (2006) and Toribio and Albert (2011) applied the posterior predictive model checking (PPMC) method, that is, available within Bayesian framework (Rubin, 1984). Theoretical advantage of PPMC method over Stone's χ^{2*} or Orlando and Thissen's $\mathrm{S}-\mathrm{X}^2$ lies in the fact that the uncertainty of item parameter estimation is taken into account in PPMC. However, simulational studies performed by the authors showed that PPMC tests were too conservative in terms of Type I error, albeit still being practically useful in terms of statistical power. Chalmers & Ng (2017) proposed a fit statistic averaged over a set of plausible values (draws from (3)) that required additional resampling to obtain the p -value. Their statistic had deflated Type I error rates, similarly to PPMC.

Problem With Existing Item-Fit Statistics

From a practical stance, the existing research on item-fit is disappointing. The researcher assessing item-fit faces a choice either to use Orlando and Thissen's $\mathrm{S}-\mathrm{X}^2$, which has low power and is not always applicable, or must refer to methods that require considerable CPU time. In consequence, they may decide not to give any consideration to statistical significance and assess item-fit merely on the value of some discrepancy measure. An example of such an approach is found in PISA

2015 technical report (OECD, 2017, p. 143), where mean deviation (MD) and root mean square deviation (RMSD) are used with disregard for their sampling properties.

The statistic proposed in this article aims at filling the gap by providing a method of testing item fit that is suitable in contexts when raw-score grouping is not applicable and is computationally feasible for practical use.

The Proposed Item-Fit Test

Let $\Delta_1, \ldots, \Delta_r$ be non-intersecting grouping intervals of ability θ, such that

$$\Delta_1 \cup \ldots \cup \Delta_r = \mathbb{R}, \quad \Delta_k \cap \Delta_h = \emptyset \text{ for } k \neq h. \tag{4}$$

The proposed approach to item-fit analysis compares two types of estimates of the expected item score over intervals Δ_k: $\widehat{\boldsymbol{O}}_j$ and $\widehat{\boldsymbol{E}}_j$. Row vector $\widehat{\boldsymbol{O}}_j$ is computed from the observed item responses, y_j, with a covariance estimate $\widehat{\boldsymbol{V}}_j$. Row vector $\widehat{\boldsymbol{E}}_j$ consists of model-based expectations that are obtained from $\widehat{f}_j$.

To test for model fit, the following Wald-type statistic is employed

$$\chi^2_{w\,j} = \left(\widehat{\boldsymbol{O}}_j - \widehat{\boldsymbol{E}}_j\right) \widehat{\boldsymbol{V}}_j^{-1} \left(\widehat{\boldsymbol{O}}_j - \widehat{\boldsymbol{E}}_j\right)^T, \tag{5}$$

which is assumed to be asymptotically chi square distributed with $r - q$ degrees of freedom, where q is the number of estimated model parameters used in computation of $\widehat{\boldsymbol{E}}_j$.

The following sections will define quantities used in equation (5) and lay out the rationale behind the asymptotic claim about χ^2_w. The presentation will be restricted to unidimensional IRT models for dichotomous items because most of the research in the field was done under these restrictions and it also allows to keep things simple. Therefore, $\boldsymbol{O}_j$ and $\boldsymbol{E}_j$ will be henceforth referred to as vectors of the pseudo-observed and expected proportions of correct responses. It should be kept in mind, however, that equation (5) is defined in terms that apply to polytomous items and equation (4) could also be defined over multidimensional ability space.

A possibility of developing a Wald-type item-fit statistic such as equation (5) was mentioned by Stone (2000), at the very end of his paper. Stone discussed $\widehat{\boldsymbol{O}}_j$ and $\widehat{\boldsymbol{E}}_j$ computed at quadrature points, rather than over ability intervals, and did not indicate any way of obtaining $\widehat{\boldsymbol{V}}_j$. To give due credit for the general idea to Stone, the symbol χ^2_w is adopted for equation (5), as in the original article.

Case When Item Parameters are Known

Assume that the IRT model holds, and the parameters of f_j are known. A posterior probability that ability of test-taker i with a response vector $\boldsymbol{y}_i$ falls into interval Δ_k is a definite integral of (3) over Δ_k

$$\tau_{ki} = g(\theta | \boldsymbol{y}_i, \Delta_k) = \int_{\Delta_k} g(\theta | \boldsymbol{y}_i) d\theta = \frac{\int_{\Delta_k} p(\boldsymbol{y}_i | \theta) g(\theta) d\theta}{p(\boldsymbol{y}_i)}. \tag{6}$$

After observing m response vectors, an estimate of O_{jk}, that is, of the pseudo-observed proportion of correct responses to item j in interval Δ_k, is given by

$$\overline{O}_{jk} = \frac{\sum_{i=1}^{m} y_{ij}\tau_{ki}}{\sum_{i=1}^{m} \tau_{ki}}. \tag{7}$$

$\overline{O}_{jk}$ is a weighted mean of item responses with weights being the posterior probabilities of test-taker membership in grouping interval Δ_k. This estimate closely resembles the ML estimate of component mean in Bernoulli mixture model (McLachlan & Peel, 2000). A mixture model analogy can be further seen by noting that for each response vector $\boldsymbol{y_i}$: $\tau_{ki} \geq 0$, $\sum\limits_{k=1}^{r} \tau_{ki} = 1$ and $p(y_j|\boldsymbol{y_i}) = \sum\limits_{k=1}^{r} \tau_{ki} p(y_j|\ \boldsymbol{y_i}, \Delta_k)$. The difference with mixture model is that *a posteriori* group membership, τ_{ki}, used in (7) is obtained "externally" from the IRT model likelihood and not estimated via the likelihood of a mixture model.

The proposed item-fit test statistic assumes that the vector of estimates of pseudo-observed proportions (7) over all ability intervals (4), $\overline{\boldsymbol{O}}_j$, is asymptotically multivariate normal with mean $\boldsymbol{O}_j$ and covariance matrix $\boldsymbol{V}_j$. That is, as $m \rightarrow \infty$

$$\left(\overline{\boldsymbol{O}}_j - \boldsymbol{O}_j\right) \boldsymbol{V}_j^{-\frac{1}{2}} \xrightarrow{d} N_r(\boldsymbol{0}, \boldsymbol{I_r}). \tag{8}$$

A test regarding $\boldsymbol{O}_j$ can be derived from (8). To verify $H_0:\ \boldsymbol{O}_j = \boldsymbol{O}_{0j}$ versus $H_1:\ \boldsymbol{O}_j \neq \boldsymbol{O}_{0j}$ the following quadratic form with an asymptotic χ_r^2 distribution is employed

$$\left(\overline{\boldsymbol{O}}_{\boldsymbol{j}} - \boldsymbol{O}_{0j}\right) \boldsymbol{V}_j^{-1} \left(\overline{\boldsymbol{O}}_{\boldsymbol{j}} - \boldsymbol{O}_{0j}\right)^T \xrightarrow{d} \chi_r^2. \tag{9}$$

The covariance matrix $\boldsymbol{V}_j$ in (9) is replaced by an estimator $\overline{\boldsymbol{V}}_j = [\overline{v}_{jkh}]_{r \times r}$, where the (k,h) th element, a covariance between $\overline{O}_{jk}$ and $\overline{O}_{jh}$, is given by

$$\overline{v}_{jkh} = \frac{\sum_{i=1}^{m} \tau_{ki}\tau_{hi}\left(y_{ij} - \overline{O}_{jk}\right)\left(y_{ij} - \overline{O}_{jh}\right)}{\left(\sum_{i=1}^{m} \tau_{ki}\right)\left(\sum_{i=1}^{m} \tau_{hi}\right)}. \tag{10}$$

As pointed out in Shao (1999, p. 404), (9) is also true if $\boldsymbol{V}_j$ is replaced by a consistent estimator.

Let $\boldsymbol{y}_{i\setminus j}$ denote a response vector of test-taker i to all items but the item j. Model-based probability of a correct response to item j in interval Δ_k upon observing $y_{i\setminus j}$ is given by

$$e_{jki} = p\left(\mathbf{y}_j = 1 \middle| \boldsymbol{y}_{i\setminus j}, \Delta_k\right) = \frac{p\left(y_j = 1, \boldsymbol{y}_{i\setminus j} \middle| \Delta_k\right)}{p\left(\boldsymbol{y}_{i\setminus j} \middle| \Delta_k\right)} = \frac{\int_{\Delta_k} f_j(\theta) p\left(\boldsymbol{y}_{i\setminus j} \middle| \theta\right) g(\theta) d\theta}{\int_{\Delta_k} p\left(\boldsymbol{y}_{i\setminus j} \middle| \theta\right) g(\theta) d\theta}. \tag{11}$$

After observing m response vectors, a model-based expected proportion of correct responses to item j in interval Δ_k to par with (7) can be computed as

$$\overline{E}_{jk} = \frac{\sum_{i=1}^{m} e_{jki}\tau_{ki}}{\sum_{i=1}^{m} \tau_{ki}}. \tag{12}$$

Finally, the item-fit is tested by stating $H_0:\ \boldsymbol{O}_j = \overline{\boldsymbol{E}}_j$ against $H_1:\ \boldsymbol{O}_j \neq \overline{\boldsymbol{E}}_j$ with a test statistic

$$\chi^2_{w\,j} = \left(\overline{\boldsymbol{O}}_j - \overline{\boldsymbol{E}}_j\right) \overline{\boldsymbol{V}}_j^{-1} \left(\overline{\boldsymbol{O}}_j - \overline{\boldsymbol{E}}_j\right)^T, \tag{13}$$

that is asymptotically chi squared with r degrees of freedom.

Case When Item Parameters are Estimated

When IRT model parameters are estimated from data, item response functions f_j are replaced with $\widehat{f}_j$; *a posteriori* group membership τ_{ki} (6) is replaced with an estimate $\widehat{\tau}_{ki}$; and pseudo-observed proportion (7), model-expected proportion (12), and covariance element (10) are replaced, respectively, by estimates

$$\widehat{O}_{jk} = \frac{\sum_{i=1}^{m} y_{ij}\widehat{\tau}_{ki}}{\sum_{i=1}^{m}\widehat{\tau}_{ki}}, \tag{14}$$

$$\widehat{E}_{jk} = \frac{\sum_{i=1}^{m} \widehat{e}_{jki}\widehat{\tau}_{ki}}{\sum_{i=1}^{m}\widehat{\tau}_{ki}}, \tag{15}$$

$$\widehat{v}_{jkh} = \frac{\sum_{i=1}^{m}\widehat{\tau}_{ki}\widehat{\tau}_{hi}\left(y_{ij} - \widehat{O}_{jk}\right)\left(y_{ij} - \widehat{O}_{jh}\right)}{\left(\sum_{i=1}^{m}\widehat{\tau}_{ki}\right)\left(\sum_{i=1}^{m}\widehat{\tau}_{hi}\right)}. \tag{16}$$

The item-fit statistic (13) for an IRT model with parameters estimated from data becomes $\chi^2_{w\,j} = (\widehat{\boldsymbol{O}}_j - \widehat{\boldsymbol{E}}_j)\,\widehat{\boldsymbol{V}}_j^{-1}(\widehat{\boldsymbol{O}}_j - \widehat{\boldsymbol{E}}_j)^T$ as previously stated in (5). Number of degrees of freedom of $\chi^2_{w\,j}$ needs to be adjusted to account for the number of estimated model parameters used in computation of $\widehat{\boldsymbol{E}}_j$.

Monte Carlo Experiments

This section describes results of three simulation studies conducted to examine properties of χ^2_w. First simulations dealt with implementation issues and their main purpose was to verify how well the asymptotic claims about χ^2_w hold upon varying approaches to construction of grouping intervals. Second simulations replicated a Monte Carlo experiment designed by Stone and Zhang (2003) that was augmented to include additional condition of incomplete response vectors. This experiment allowed to analyze Type I error rates and power of χ^2_w against a benchmark of Orlando and Thissen's $\mathrm{S}-\mathrm{X}^2$ and Stone's χ^{2*}, and to verify performance of χ^2_w in an incomplete-data design setting. Final study was based on the "bad items" design by Orlando and Thissen (2003) and aimed at providing further information about power of χ^2_w.

Ability parameters in all three simulation studies were sampled from normal distribution $g(\theta) = N(0,1)$. Item response functions belonged to the logistic family of IRT models: the three-parameter logistic model (3PLM)

$$f_j(\theta) = P\left(y_j|\theta\right) = c_j + \frac{1 - c_j}{1 + e^{-a_j\left(\theta - b_j\right)}}, \tag{17}$$

the two-parameter logistic model (2PLM, (17) with $c_j = 0$) and the one-parameter logistic model (1PLM, (17) with $c_j = 0$ and $a_j = a$).

All analyses were performed in Stata. Item responses under (17) were generated using uirt_sim (Kondratek, 2020). Parameters of IRT models were estimated by uirt (version 2.1; Kondratek, 2016) with its default settings—EM algorithm, Gauss–Hermite quadrature with 51 integration points, and 0.0001 stopping rule for maximum absolute change in parameter values between EM iterations. The uirt software was also used to compute $\text{S}-\text{X}^2$ and χ^2_w. Indefinite integrals over $g(\theta)$, needed for expected proportions used in $\text{S}-\text{X}^2$ and to obtain $p(\boldsymbol{y_i})$ seen in denominator of (6), were computed by Gauss–Hermite quadrature with 151 integration points. Definite integrals over $g(\theta)$, seen in the numerator of (6), employed Gauss–Legendre quadrature with 30 integration points at each bin of ability, Δ_k.

Each of the Monte Carlo experiments involved 10,000 replications of the simulated conditions. Type I error and power were computed as percentage of rejected H_0 at significance level $\alpha = 0.05$, averaged over replications.

Simulation Study 1 – Number and Range of Grouping Intervals

Implementation of χ^2_w has required decisions on the number of grouping intervals (4) and their range. Postulated distribution of χ^2_w is derived from asymptotic normality of the vector of the pseudo-observed proportions (8); therefore, the conventional rule for appropriateness of normal approximation to sample proportion was adopted to govern range of ability intervals. This resulted in item-specific intervals Δ_k that were constructed so that

$$m_{jk}\widehat{\pi}_{jk}\left(1-\widehat{\pi}_{jk}\right) \approx const, \tag{18}$$

where $\widehat{\pi}_{jk}$ is a simple model-based estimate of proportion of correct responses in interval Δ_{jk}

$$\widehat{\pi}_{jk} = \widehat{p}\left(y_j = 1|\Delta_{jk}\right) = \frac{\int_{\Delta_{jk}} \widehat{f}_j(\theta)g(\theta)d\theta}{\int_{\Delta_{jk}} g(\theta)d\theta}. \tag{19}$$

and m_{jk} is the expected number of observations in interval Δ_{jk}, $m_{jk} = m\int_{\Delta_{jk}} g(\theta)d\theta$.

Ranges of Δ_{jk} that would meet the condition (18) were determined by first splitting the ability distribution into smaller Δ_{jv} intervals that were equiprobable with respect to $g(\theta)$, $m_{jv} = 0.001m$, and $v\in\{1,\ldots,1000\}$. The finer intervals were then aggregated into $\Delta_{jk} = \text{U}_{v=a}^{v=b}\Delta_{jv}$ so that $\sum\limits_{v=a}^{v=b}\widehat{\pi}_{jv}(1-\widehat{\pi}_{jv}) \approx \frac{1}{r}\sum\limits_{v=1}^{v=1000}\widehat{\pi}_{jv}(1-\widehat{\pi}_{jv})$, where r was the desired number of Δ_{jk} (4). Computation of $\widehat{\pi}_{jv}$ in this step was performed with Gauss–Legendre quadrature with 11 integration points.

Behavior of χ^2_w upon adopting criterion (18) with three ability bins when testing fit of an easy 2PLM item ($a_j = 1.7$ and $b_j = -1.84$) under true H_0 is illustrated in Figure 1 (upper panel) and compared to an alternative equiprobable division of ability (lower panel). Graphs presented in Figure 1 were obtained in a simple Monte Carlo experiment in which the item tested for fit was embedded in a 30-item 2PLM test. The remaining 29-item parameters were sampled from $\ln a_{v\neq j} \sim N(\ln 1.7, 0.4)$ and $b_{v\neq j} \sim N(0,1)$, and the sample size was $m = 1000$. In each replication, $\chi^2_{w\,j}$ was computed and the pseudo-observed proportions of correct responses $\widehat{O}_{jk}$ (14) were stored. Upon completing 10,000 replications, the resulting empirical distribution of $\chi^2_{w\,j}$ was compared against theoretical $\chi^2(1)$ in a Q–Q plot, and the pseudo-observed proportions were transformed according to (8) so that standardized variables were obtained and compared against theoretical $N(0,1)$ on histograms.

The equiprobable division would be a tempting alternative to (18) as it results in equal expected number of observations in each interval, $m_{jk} \approx const$. By being independent from item parameters,

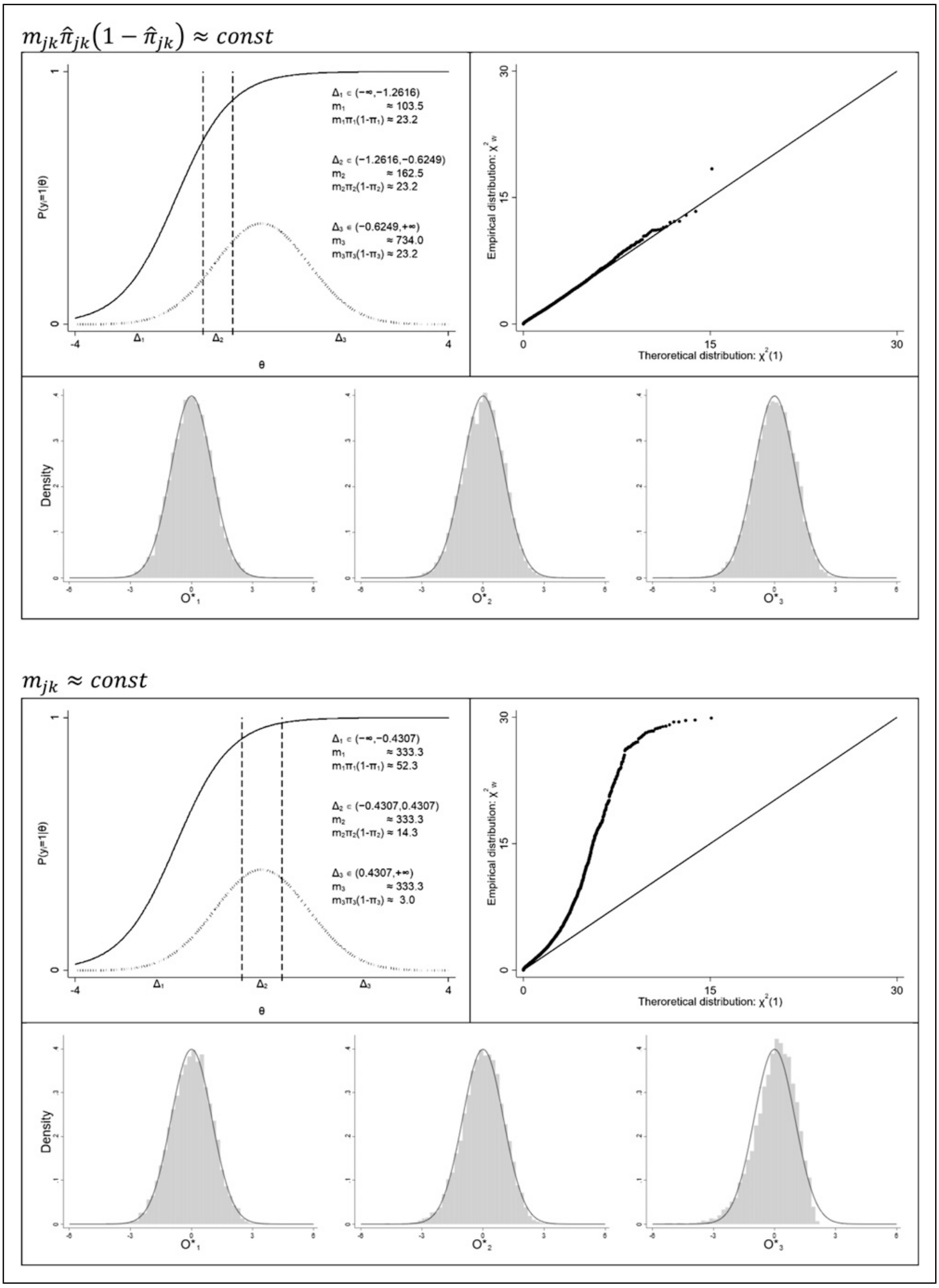

Figure 1. Distribution of χ^2_w under different choices of interval range.

it would decrease the computational cost of χ^2_w because the group membership probabilities, $\widehat{\tau}_{ki}$, would need to be obtained only once. However, simulation results presented in Figure 1 indicate that $\chi^2_{w\,j}$ grossly deviated from theoretical $\chi^2(1)$ upon such division. The expected number of observations in each bin exceeds 333, but because of the extreme easiness of the item, the rightmost bin is associated with a very small value on the $m_{jk}\widehat{\pi}_{jk}(1-\widehat{\pi}_{jk})$ criterion. The transformed proportion of correct responses in this bin experiences a visible ceiling effect, and thus the $\chi^2(1)$ assumption does not hold. Yet, when $\chi^2_{w\,j}$ was computed with intervals constructed using criterion (18), it resulted in a good approximation to $\chi^2(1)$, even at these rather difficult conditions in terms of sample size and item difficulty.

A second problem of implementation of χ^2_w required deciding on the number of ability bins, r. It was expected that increasing r would be detrimental to the normal approximation of pseudo-observed proportions. So, from the standpoint of Type I error, the safest approach would be to use the smallest possible number of intervals, $r = q + 1$, leading to χ^2_w with a single degree of freedom. However, the relation between r and power of χ^2_w was not obvious. On the one hand, increasing r would allow to detect a locally finer grade of deviances between (14) and (15). On the other hand, it would increase the entries of the covariance matrix (16) because of smaller effective sample size per interval.

Number of Ability Bins – Simulation Design. To investigate properties of χ^2_w under varying number of bins, a Monte Carlo experiment was conducted under similar scheme that was used to obtain results reported in Figure 1. The conditions of experiment were extended to cover different IRT models (1PLM, 2PLM, and 3PLM), items of varying marginal difficulty, $\pi_j = \int f_j(\theta)g(\theta)d\theta$ ($\pi_j = 0.5$ and $\pi_j = 0.9$ for all models and additionally $\pi_j = 0.3$ for 3PLM), two sample sizes ($m \in \{400,\ 4000\}$), and two test lengths ($n \in \{10,\ \ 40\}$). Under each of these conditions Type I error was obtained in both fixed and estimated parameters case, and Q–Q plots were plotted for certain number of ability bins for closer assessment of the distribution of χ^2_w. Additionally, power to detect misfit was analyzed with varying number of ability bins by fitting a 1PLM or 2PLM model to an item estimated under 3PLM. This article presents only main conclusions from the experiment; detailed results under all tested conditions are provided in the online supplement.

Number of Ability Bins – Simulation Results. Figure 2 depicts relationship between the number of ability intervals and resulting detection rates for an item of medium difficulty ($\pi_j = 0.5$) that was simulated as 3PLM and then estimated as 3PLM (Type I error) or as either 2PLM or 1PLM (statistical power). It illustrates that increasing the number of ability bins results in decrease of statistical power of χ^2_w. Additionally, increase of r eventually leads to elevated Type I error rates. These patterns were also seen in other conditions considered in the experiment, with the detrimental effect of increased r on the Type I error being especially prominent for difficult items in small samples (online supplement).

Based on these results, it was decided to implement χ^2_w with $r = q + 1$ intervals for 3PLM and 2PLM. For 1PLM: either $r = 3$ if criterion (18) exceeds 20, or $r = 2$ otherwise. Number 20 precautiously doubles the conventional rule for when a normal approximation to sample proportion is appropriate. These settings were used in all the simulation studies covered in the rest of the article.

Q–Q plots were obtained according to the adopted rule for the number of intervals for a more detailed verification of the postulated asymptotic distribution of χ^2_w (online supplement). For $m = 4000$, the empirical distribution of χ^2_w was well aligned with theoretical χ^2 under all tested conditions, both in the known and the estimated parameters case. Approximation was also well-behaved for $m = 400$ and moderate item difficulty. However, combined conditions of small sample size and extreme item difficulties resulted in deviation of χ^2_w from its theoretical asymptotic distribution. We should notice that under such conditions, the criterion for appropriateness of normal approximation of sample proportion (18) is small in value, even when the lowest possible number of ability intervals is used. This alerts us that χ^2_w should be used with caution whenever fit of extremely easy or difficult item is to be assessed in small samples. A condition $m_{jk}\widehat{\pi}_{jk}(1 - \widehat{\pi}_{jk}) > 20$ seems to be a good guideline on deciding if results of χ^2_w are trustworthy (see Figure 1).

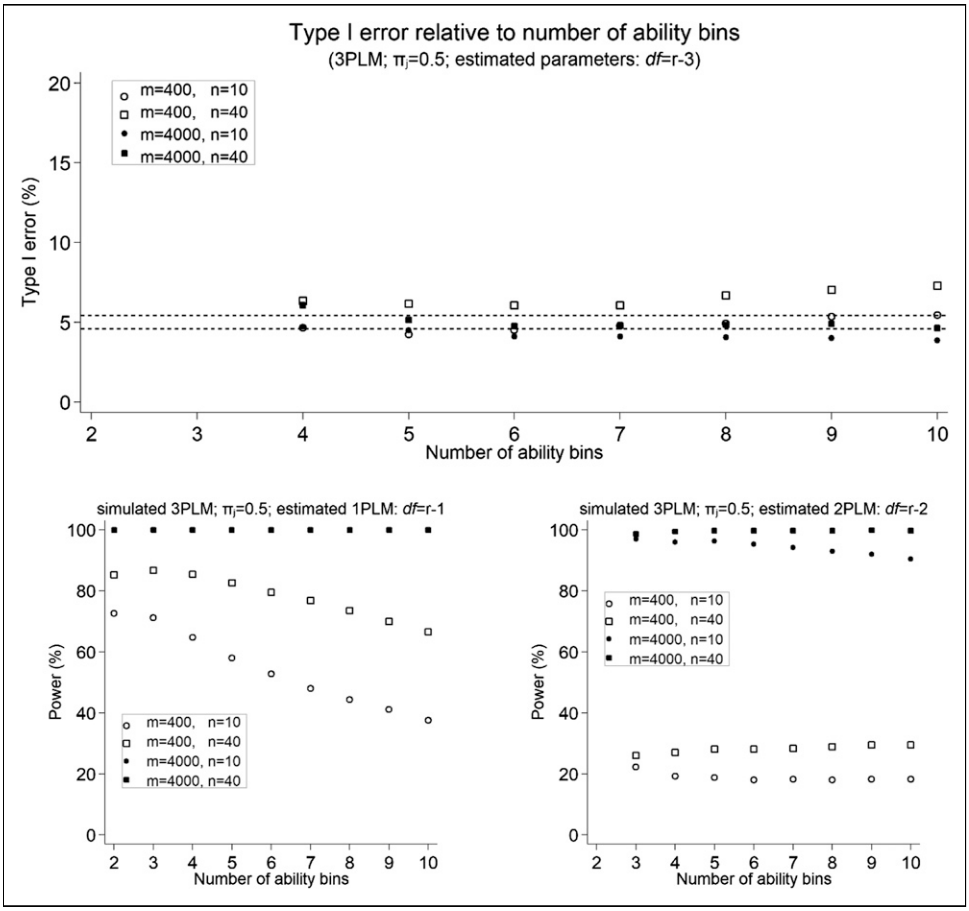

Figure 2. Type I error and power of χ^2_w in relation to the number of ability bins.

Simulation Study 2 – Type I Error and Power

Simulation Design. This study replicated Monte Carlo experiment designed by Stone and Zhang (2003). Three test lengths, $n \in \{10, 20, 40\}$, were crossed with three sample sizes, $m \in \{500, 1000, 2000\}$, and data was generated under two IRT models: 2PLM and 3PLM. Under the 2PLM generating scenario, a set of 10 pairs of item parameters was constructed by crossing two values of item discrimination parameter, $a_j \in \{1.2, 2.2\}$, with five values of item difficulty parameter, $b_j \in \{-2, -1, 0, 1, 2\}$. 20-item and 40-item tests were built by adding another 10 or 3x10 items defined by repetition of the same parameter set. The 3PLM scenario used the same discriminations and difficulties as the 2PLM. All items, except the easiest one with $b_j = -2$, were added a pseudo-guessing parameter $c_j = 0.25$ in the 3PLM scenario.

This design was extended to create additional incomplete-data conditions. It was accomplished by taking the complete data generated under the original design for $n \in \{20, 40\}$ and $m \in \{1000, 2000\}$, and treating random 50% of responses for each item as missing. In result, additional four generating conditions were introduced in which number of observations per item and expected number of items per observation halved the size of the original complete data.

Table 1. Type I error rates for different item-fit statistics (%).

Test length	Sample size	Results of current study			Results of Stone and Zhang	
		$S-X^2$ Complete data	χ^2_w Complete data	χ^2_w 50% missing	$S-X^2$	χ^{2*}
$n = 10$	$m = 500$	4.8	4.2		4	5
	$m = 1000$	4.7	3.8		4	4
	$m = 2000$	4.9	3.7		5	3
$n = 20$	$m = 500$	4.9	4.7		5	5
	$m = 1000$	4.9	4.3	4.3	5	3
	$m = 2000$	5.0	4.2	4.1	5	3
$n = 40$	$m = 500$	4.7	5.1		6	6
	$m = 1000$	4.8	4.8	4.8	4	4
	$m = 2000$	5.0	4.7	4.5	3	4

Note. Stone and Zhang (2003, Table 1).

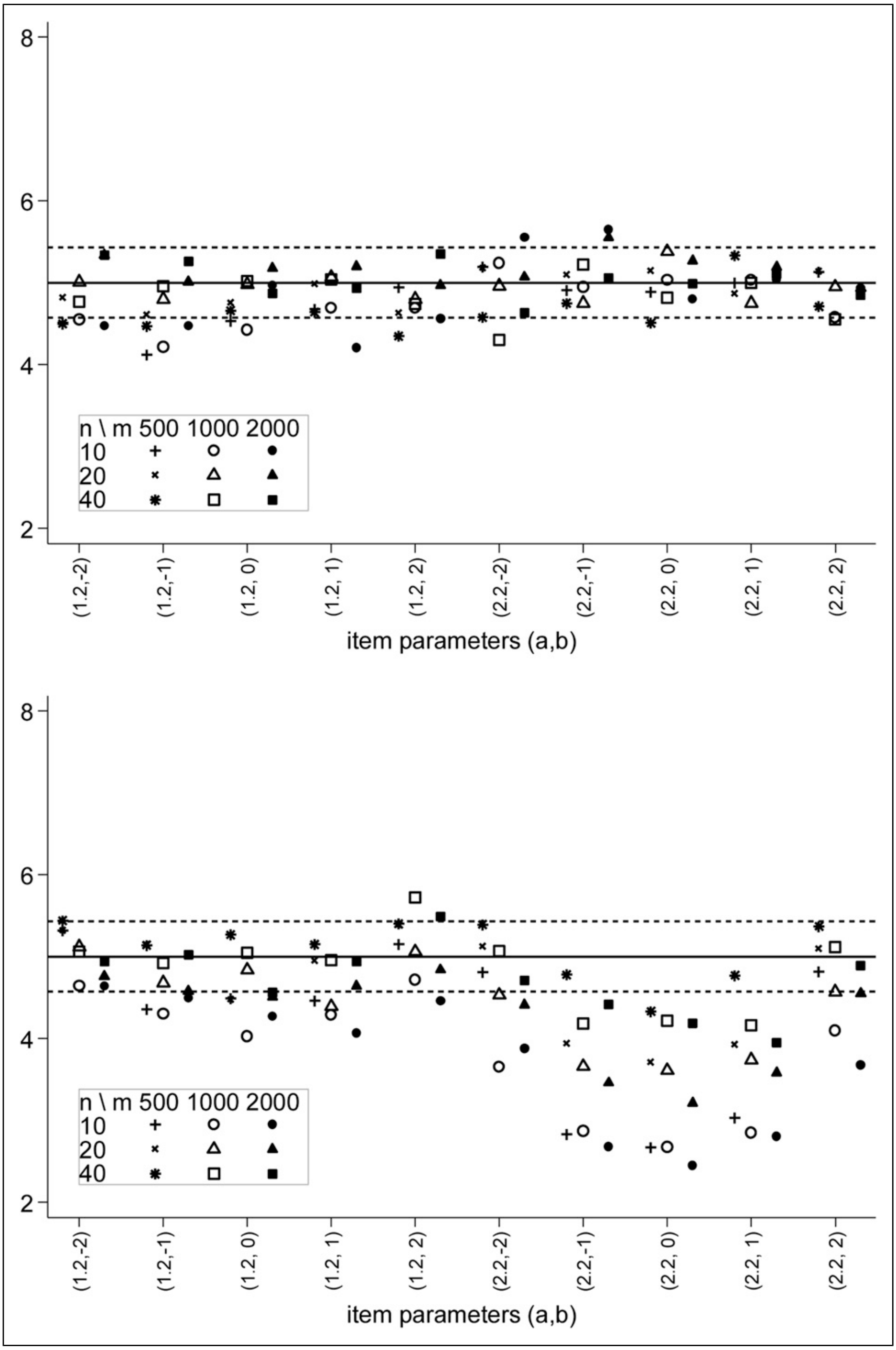

Figure 3. Type I error rates for $S - X^2$ (top) and χ_w^2 (bottom) conditional on item parameters.

Table 2. Power rates for different item-fit statistics (%).

Test length	Sample size	Results of current study			Results of Stone and Zhang	
		$S-X^2$ Complete data	χ^2_w Complete data	χ^2_w 50% missing	$S-X^2$	χ^{2*}
Simulated 2PLM – Estimated 1PLM						
$n=10$	$m=500$	23.8	35.1		26	51
	$m=1000$	49.5	59.9		53	75
	$m=2000$	80.7	86.5		81	94
$n=20$	$m=500$	23.2	47.5		23	56
	$m=1000$	46.2	73.1	36.4	45	78
	$m=2000$	78.7	93.7	60.8	80	96
$n=40$	$m=500$	20.3	52.3		22	52
	$m=1000$	38.4	77.2	46.6	40	78
	$m=2000$	71.9	95.5	71.9	75	94
Simulated 3PLM – Estimated 1PLM						
$n=10$	$m=500$	36.5	47.8		35	69
	$m=1000$	58.6	64.3		59	82
	$m=2000$	75.0	77.8		75	88
$n=20$	$m=500$	41.4	61.9		42	67
	$m=1000$	63.7	74.9	49.7	65	80
	$m=2000$	75.9	86.2	66.0	77	87
$n=40$	$m=500$	41.5	66.2		42	68
	$m=1000$	63.4	78.0	61.5	64	80
	$m=2000$	74.6	88.2	74.6	75	90
Simulated 3PLM – Estimated 2PLM						
$n=10$	$m=500$	7.1	23.7		7	13
	$m=1000$	9.3	40.2		10	30
	$m=2000$	13.3	62.4		14	46
$n=20$	$m=500$	8.3	25.9		9	13
	$m=1000$	11.5	41.0	23.9	10	25
	$m=2000$	16.8	58.6	40.2	17	44
$n=40$	$m=500$	8.0	20.6		6	15
	$m=1000$	11.4	32.1	25.1	7	28
	$m=2000$	17.4	46.1	40.3	12	44

Note. Stone and Zhang (2003, Table 2).

In each replication, all item parameters were estimated from the simulated data under two IRT models: 1PLM and 2PLM. In complete data conditions, Orlando and Thissen's $\mathrm{S}-\mathrm{X}^2$ and the χ^2_w statistics were computed from the estimates of the IRT model. In the missing responses scenario, only χ^2_w was obtained because $\mathrm{S}-\mathrm{X}^2$ is not applicable to data with incomparable sum-scores. The case when generating model was the same as estimating model (2PLM) served to analyze Type I error. Other three combinations, when generating model had more parameters than estimating model, were used to assess statistical power of $\mathrm{S}-\mathrm{X}^2$ and χ^2_w.

Simulation Results. Table 1 summarizes performance of $\mathrm{S}-\mathrm{X}^2$ and χ^2_w when both the generating and the estimating model were 2PLM. Entries in the table are percentages of rejected H_0 at significance level $\alpha = 0.05$ averaged over all items and all replications. Results for $\mathrm{S}-\mathrm{X}^2$ and χ^{2*} reported in Stone & Zhang (2003) are also included for reference. It should be kept in mind that Stone and Zhang results were obtained with two orders of magnitude fewer replications. Figure 3 expands the analysis of Type I errors of $\mathrm{S}-\mathrm{X}^2$ and χ^2_w, by presenting rejection rates of true H_0 at an item level. Results for the first 10 items are plotted against a 95% confidence bound around the nominal significance level $\alpha = 0.05$, assuming standard error of $\sqrt{\alpha(1-\alpha)/10^4}$.

Type I error of $\mathrm{S}-\mathrm{X}^2$, as seen in Table 1 and Figure 3, was almost flawlessly nominal for all items, test-lengths, and sample sizes that were considered in the study. This result confirms what was previously observed by Orlando and Thissen (2000) or Stone and Zhang (2003).

The averaged Type I error of χ^2_w was in 0.037–0.045 range (Table 1) which is acceptable. These values do not exceed ranges reported for both $\mathrm{S}-\mathrm{X}^2$ and χ^{2*} by Stone and Zhang (2003) under the same experimental conditions. However, the item-level information in Figure 3 reveals that for higher discriminating items ($a_j = 2.2$) with moderate difficulty ($b_j \in \{-1, 0, 1\}$) Type I error of χ^2_w is deflated. This effect diminishes with increase in test length. From practical standpoint deflated false rejection rates would not be a problem as long as χ^2_w has sufficient power to detect misfit.

Power of $\mathrm{S}-\mathrm{X}^2$ and χ^2_w was examined by averaging rejection rates in three scenarios when the model used for simulating responses had more item parameters than the model used in estimation: 2PLM-1PLM, 3PLM-1PLM, and 3PLM-2PLM. The results are presented in Table 2, together with power of $\mathrm{S}-\mathrm{X}^2$ and χ^{2*} from simulation by Stone and Zhang (2003). The χ^2_w statistic outperformed $\mathrm{S}-\mathrm{X}^2$ with regard to power under all experimental conditions. When compared to results for χ^{2*} reported in Stone and Zhang (2003), χ^2_w was more sensitive in detecting misfit under the 3PLM-2PLM. In other misfit scenarios χ^{2*} and χ^2_w achieved similar power under long tests ($n = 40$) and the reported power of χ^{2*} exceeded that of χ^2_w for shorter tests. Power of Stone's χ^{2*} seems to be unaffected by the test length. This puts Stone's χ^{2*} in a position of an especially useful item-fit measure for short tests.

To conclude remarks on this Monte Carlo experiment, it is worth noticing that χ^2_w performed well when random 50% of item responses were missing both in terms of its averaged Type I error rates (Table 1) and power (Table 2). Results for χ^2_w under the missing responses condition closely

Table 3. Power rates for three types of misfitting items.

Item	Test length	*n* = 500		*n* = 1000		*n* = 2000	
		$\mathrm{S}-\mathrm{X}^2$	χ^2_w	$\mathrm{S}-\mathrm{X}^2$	χ^2_w	$\mathrm{S}-\mathrm{X}^2$	χ^2_w
BAD1	m = 10	0.379	0.634	0.551	0.845	0.790	0.969
	m = 20	0.539	0.868	0.753	0.982	0.955	0.999
	m = 40	0.598	0.964	0.861	0.999	0.992	1.000
	m = 80	0.533	0.987	0.865	1.000	0.998	1.000
BAD2	m = 10	0.130	0.228	0.209	0.413	0.378	0.680
	m = 20	0.221	0.450	0.406	0.749	0.731	0.953
	m = 40	0.315	0.607	0.586	0.875	0.912	0.993
	m = 80	0.351	0.641	0.659	0.905	0.957	0.996
BAD3	m = 10	0.221	0.520	0.444	0.802	0.783	0.969
	m = 20	0.359	0.764	0.756	0.967	0.982	1.000
	m = 40	0.443	0.842	0.873	0.990	1.000	1.000
	m = 80	0.444	0.851	0.878	0.992	1.000	1.000

resemble the ones that are observed for complete data but with twice less items and observations, which is exactly the expected outcome. This result puts χ^2_w at advantage of over methods that rely on observed sum-score portioning of ability, like $\mathrm{S}-\mathrm{X}^2$.

Simulation Study 3 – Power

Simulation Design. Last experiment adapted a design proposed by Orlando and Thissen (2003) to analyze power in misfit scenarios that go beyond fitting of a restricted IRT model to data generated from an unrestricted model. It involved three "bad" items that are described by response functions

$$\mathrm{BAD1}:\ P(y_j|\theta) = \frac{c_j}{1+e^{a_j(\theta-(b_j-d_j))}} + \frac{1}{1+e^{-a_j(\theta-b_j)}},$$

where $a_j = 1.7, 2.5$, $b_j = 1$, $c_j = 0.25$, and $d_j = 1.5$;

$$\mathrm{BAD2}:\ P(y_j|\theta) = \frac{d_j}{1+e^{-a_j(\theta-b_j)}},$$

where $a_j = 1.7, 2$, $b_j = 0.5$, $d_j = 0.7$; and

$$\mathrm{BAD3}:\ P(y_j|\theta) = \frac{x_j}{1+e^{-a_j(\theta-b_j)}} + \frac{y_j}{1+e^{a_j(\theta-(b_j-d_j))}},$$

where $a_j = 1.7, 3.5$, $b_j = -1$, $d_j = 3$, $x_j = 0.55$, and $y_j = 0.45$.

These bad items were, one at a time, embedded in tests consisting of $n \in \{10, 20, 40, 80\}$ total items. The remaining $v \neq j$ items were drawn from 2PLM with $\ln a_{v\neq j} \sim N(0, 0.5)$ and $b_{v\neq j} \sim N(0, 1)$. For each test length $m \in \{500, 1000, 2000\}$, item responses were generated and IRT model was fit to data. Items BAD2, BAD3, and all $v \neq j$ items were modeled with 2PLM without imposing priors on item parameters, and item BAD1 was modeled with 3PLM using noninformative priors: $N(0, 3)$ for b_j, $N(1.1,3)$ for a_j, and $\beta(1.01,1.03)$ for c_j. Estimated item parameters were used to compute χ^2_w and $\mathrm{S}-\mathrm{X}^2$ for the three bad items.

This design deviated from the original conditions used by Orlando and Thissen (2003) by adopting 3PLM only for the item BAD1, instead of using it for all items. It was motivated by observation that the c_j parameter for items BAD2 and BAD3 approached 0 with increase of m. The 3PLM would be an unnecessarily over-parametrized choice for items BAD2 and BAD3.

Simulation Results. Resulting power rates (Table 3) support previous evidence (Table 2) that χ^2_w is more sensitive in detecting misfit than $\mathrm{S}-\mathrm{X}^2$. Power of both statistics rose with increase of test length and sample size, but under all tested conditions χ^2_w exceeded $\mathrm{S}-\mathrm{X}^2$.

Summary

Multiple Monte Carlo experiments were conducted to examine properties of the new χ^2_w item-fit statistic. Type I error of χ^2_w was close to nominal level. It outperformed Orlando and Thissen's $\mathrm{S}-\mathrm{X}^2$ on power under all tested conditions. In the 3PLM-2PLM, misfit scenario χ^2_w was also more sensitive in comparison with Stone's χ^{2*}. The results are promising and χ^2_w poses as a viable candidate to test for item fit. It is especially attractive because it can be applied to incomplete testing designs, unlike alternatives that use observed scores for partitioning, and is far less computationally demanding than available statistics that involve residuals over the latent trait.

It is worth pointing to the possibility of other applications of the item-fit approach that was proposed in the article. First, χ^2_w is straightforwardly generalizable to polytomous items and to multivariate abilities. Also, the quadrature used in implementation of χ^2_w can be replaced with other solutions to cover cases when ability is not normally distributed. Second, estimates of observed proportions and of the covariance matrix in (5) can be utilized to construct confidence bounds around observed proportions. Such confidence intervals can be plotted against $\widehat{f}_j$ to aid graphical analysis of item-fit. And finally, approach outlined in the article can also be applied to perform differential item functioning (DIF) analysis.

It should be noted that mathematical underpinnings of χ_w^2 laid out in the article are incomplete. Asymptotic multivariate normality of vector of pseudo-observed proportions, (8), is assumed without proof. Consistency of the proposed estimator of the covariance, (10), is likewise just assumed. Careful consideration should also be exercised on how replacing item response functions in the known parameter case of χ_w^2 by their ML estimates impacts the asymptotic claims about χ_w^2 – especially when item parameters are estimated with priors. Results of simulational studies support asymptotic claims made about χ_w^2. However, they cannot be automatically generalized to cover conditions that would deviate from the specific ones that were considered here. This opens ground for future research on χ_w^2.

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 disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Centre research grant number 2015/17/N/HS6/02965.

ORCID iD

Bartosz Kondratek https://orcid.org/0000-0002-4779-0471

Supplemental Material

Supplemental material for this article is available online.

References

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

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/bf02291180

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*(2), 141–162. https://doi.org/10.1111/jedm.12039

Chalmers, R. P., & Ng, V. (2017). Plausible-Value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, *41*(5), 372–387. https://doi.org/10.1177/0146621617692079

Chon, K. W., Lee, W., & 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

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, *64*(3), 273–294. https://doi.org/10.1007/bf02294296

Glas, C. A. W., & Suarez-Falcon, 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., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*(4), 635–659. https://doi.org/10.1007/bf02296401

Kondratek, B. (2016). uirt: Stata module to fit unidimensional Item Response Theory models. In *Statistical Software Components S458247*. Boston College Department of Economics.

Kondratek, B. (2020). uirt_sim: Stata module to simulate data from unidimensional Item Response Theory models. In *Statistical software components S458749*. Boston College Department of Economics.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49–57. https://doi.org/10.1177/014662168500900105

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Willey.

Organization for Economic Co-operation and Development (OECD) (2017). *PISA 2015 technical report*. Paris, France: OECD.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Shao, J. (1999). *Mathematical statistics*. New York: Springer-Verlag.

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

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., & Hansen, M.A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement*, *60*(6), 974–991. https://doi.org/10.1177/00131640021970907

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

Swaminathan, H., Hambleton, R. K., & Rogers, H.J. (2007). Assessing the fit of item response theory models. In C. R. Rao, & S. Sinharay (Eds), *Handbook of statistics*. New York, NY: Elsevier.

Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for item fit analysis in item response theory. *Journal of Statistical Computation and Simulation*, *81*(10), 1345–1360. https://doi.org/10.1080/00949655.2010.485131

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*(4), 339–368. https://doi.org/10.2307/1165054

Woods, C. M. (2008). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applied Psychological Measurement*, *32*(5), 371–384. https://doi.org/10.1177/0146621607307691

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