JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12207 *Journal of Educational Measurement*

Performance of Person-Fit Statistics Under Model Misspecification

Seong Eun Hong and Scott Monroe
University of Massachusetts Amherst
Carl F. Falk
McGill University

In educational and psychological measurement, a person-fit statistic (PFS) is designed to identify aberrant response patterns. For parametric PFSs, valid inference depends on several assumptions, one of which is that the item response theory (IRT) model is correctly specified. Previous studies have used empirical data sets to explore the effects of model misspecification on PFSs. We further this line of research by using a simulation study, which allows us to explore issues that may be of interest to practitioners. Results show that, depending on the generating and analysis item models, Type I error rates at fixed values of the latent variable may be greatly inflated, even when the aggregate rates are relatively accurate. Results also show that misspecification is most likely to affect PFSs for examinees with extreme latent variable scores. Two empirical data analyses are used to illustrate the importance of model specification.

In educational and psychological measurement, a person-fit statistic (PFS) is generally designed to detect aberrant patterns of item scores (Meijer & Sijtsma, 2001), and such statistics may aid researchers in learning about an individual's answering behavior. For example, PFSs may be used by researchers to explore whether an examinee lacks motivation (Conijn, Emons, & Sijtsma, 2014), or has cheated on a test (Cizek & Wollack, 2017). Though numerous nonparametric PFSs have been developed (Karabatos, 2003), the current research focuses on statistics based on item response theory (IRT). In particular, this research studies the class of asymptotically correct PFSs of Snijders (2001), denoted as S^*.

From a statistical perspective, satisfactory performance of S^* under the null hypothesis of non-aberrant responding depends on several assumptions, each of which has been studied in the literature. One assumption is that the true item parameters, $\boldsymbol{\beta}_0$, are known (Snijders, 2001). In some practical settings, such as operational testing with large calibration samples, the variability of the item parameter estimates, $\hat{\boldsymbol{\beta}}$, will typically be small. In such cases, the use of $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}_0$ should not greatly affect S^*. For sample sizes of at least 1,000 examinees, Molenaar and Hoijtink (1990) and Glas and Dagohoy (2007) found that use of $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}_0$ did not greatly affect the performance of various PFSs. Consequently, researchers routinely use $\hat{\boldsymbol{\beta}}$ when calculating S^* (de la Torre & Deng, 2008; Sinharay, 2016a; Van Krimpen-Stoop & Meijer, 1999).

A second assumption is that the number of items is sufficiently large for S^* to follow its asymptotic distribution. Snijders (2001) proved that for long tests, the asymptotic null distribution of S^* is standard normal. However, for some realistic

test lengths (e.g., 30 items), researchers have found that the Type I error rates of S^* at significance levels of .01 or .02 are not well calibrated (de la Torre & Deng, 2008; Snijders, 2001).[1] To address this issue, de la Torre and Deng (2008) proposed the use of a person-specific null-distribution based on resampling, an approach generalized in Sinharay (2016b). Researchers have found that null distributions based on resampling can lead to more precise Type I error rates at small significance levels, especially for shorter tests (de la Torre & Deng, 2008).

A third assumption, and the focus of the current research, is that the IRT model is correctly specified. In an early study, Drasgow (1982) argued that model specification is one of the most important considerations when assessing person fit. Drasgow (1982) examined the effects of fitting either a three-parameter logistic (3PL) or one-parameter logistic (1PL) model on detection of aberrant examinees. Drasgow (1982) simulated aberrant response patterns and embedded them within real data from the Verbal section of the Graduate Record Examination (GRE). Focusing on detection of these aberrant response patterns, the author concluded that the choice of IRT model did not greatly affect the performance of the studied PFS, although the 3PL led to slightly better detection rates. In another study, Meijer and Tendeiro (2012) recommended examination of model fit as an important preliminary step in person-fit analysis. Analyzing an empirical data set, the authors used the $S - X^2$ (Orlando & Thissen, 2000) item fit statistic, as well as the standardized local-dependence statistic, X^2_{LD} (Chen & Thissen, 1997), to conclude that the 2PL model provided acceptable fit. The authors also fit the 1PL model to the data to illustrate the effect of model misspecification, and they concluded that the greater number of flagged patterns under the 1PL model was due to model misspecification. Importantly, both Drasgow (1982) and Meijer and Tendeiro (2012) based their studies on empirical data, and focused primarily on aggregate rejection rates, as opposed to rates conditional on ability.

The current research extends these earlier studies by further exploring the effect of IRT model specification on the performance of PFSs. Though model fit is routinely examined as part of item analysis in operational testing programs, we argue the effect of model specification on PFSs nevertheless deserves attention, for several reasons. First, items that exhibit misfit may nevertheless be retained due to practical considerations. For example, such items may be retained to balance content coverage or to satisfy test assembly needs (Crisan, Tendeiro, & Meijer, 2017). As another example, misfitting items may be retained because test developers typically choose a single model for all items of a certain type (e.g., 2PL for all multiple-choice items) to facilitate communication with stakeholders and preserve compatibility with existing operational infrastructure (Zhao & Hambleton, 2017). Second, model fit evaluation is a complex endeavor (Maydeu-Olivares, 2013) that may likewise be guided to some degree by practical concerns. Consequently, there is no consensus regarding how model fit should be evaluated. Finally, the fit of the model may be evaluated only with respect to the primary intended use, which is almost certainly not person-fit assessment. That is, the primary intended use (e.g., proficiency classification) may be quite robust to certain misspecifications, while a secondary intended use (e.g., producing PFSs) may be quite sensitive to the same misspecification (Sinharay & Haberman, 2014). For all of these reasons, it should not be assumed in operational

settings that PFSs are unaffected by model misspecification, despite a preliminary model fit evaluation.

In contrast to earlier research, we use a simulation study to explore how PFSs are affected by model misspecification. This approach allows us to focus on Type I error rates conditional on a fixed value of the latent achievement. That is, we can examine whether model misspecification equally affects examinees at different levels of the latent trait. Though model misspecification may take countless forms, this research focuses on the most common logistic item models (i.e., 1PL, 2PL, and 3PL) for dichotomously scored items. The primary contribution of the current research, then, is to conduct such a simulation study, with the goal of providing valuable insights to researchers who use PFSs in practice. In brief, Type I error rates of popular variants of S^* are compared when data are generated under the 1PL, 2PL, or 3PL, but when the fitted model may or may not be correct. As a secondary contribution, a graphical approach for studying conditional rejection rates of PFSs with empirical data is illustrated. Based on results from the simulation study and empirical data analyses, recommendations are provided on how practitioners may assess the impact of misspecification on PFSs. The remainder of this paper reviews the methodology used in this study, presents the simulation design and results, and provides two real data examples, before offering some concluding remarks.

Methodological Background

The purpose of this section is to present how model misspecification affects calculation of S^* (Snijders, 2001). The presentation is limited to key ideas, and technical details are avoided. Interested readers are referred to Snijders (2001) and Magis, Raich, and Beland (2012).

The PFS of Snijders (2001) Under Correct Model Specification

Let there be $i = 1, \ldots, N$ examinees and $j = 1, \ldots, n$ dichotomous items. Let the ith response pattern be $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in})$, where the item response y_{ij} is either 0 or 1. Let the true item response function (IRF) be

$$P_j(\theta_i) = \Pr(y_{ij} = 1 | \theta_i, \boldsymbol{\beta}_0; \mathcal{M}_0), \tag{1}$$

where θ_i is the latent ability, $\boldsymbol{\beta}_0$ is the collection of true item parameters, and $\mathcal{M}_0$ denotes the true item model. For example, if $\mathcal{M}_0$ is the 3PL, then for item j,

$$P_j(\theta_i) = g_{0j} + (1 - g_{0j}) \frac{1}{1 + \exp[-a_{0j}(\theta_i - b_{0j})]}, \tag{2}$$

where a_{0j}, b_{0j}, and g_{0j} are the true discrimination, difficulty, and guessing parameters, respectively. The 2PL and 1PL models may be obtained as special cases of Equation 2. Given the availability of a calibration sample, the maximum likelihood (ML) estimate of $\boldsymbol{\beta}_0$, say $\hat{\boldsymbol{\beta}}$, may be obtained. Assuming correct model specification, estimates will converge to the true values, $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$, as sample size increases.

Snijders (2001) derived the asymptotic null distribution for a class of PFSs, assuming the model is correctly specified and $\boldsymbol{\beta}_0$ is known. Under these assumptions, let $\hat{\theta}_i$ be an ability estimate based on the response pattern $\boldsymbol{y}_i$ and item parameters

$\boldsymbol{\beta}_0$, such as the "maximum a posteriori" (MAP) or the weighted likelihood estimate (WLE; Warm, 1989). Note that $\hat{\theta}_i$ depends on the true IRFs. For several popular ability estimates, including the MAP and WLE estimates, Snijders (2001) showed that a class of PFSs, which may be written as

$$S^*(\hat{\theta}_i) = S^*(\boldsymbol{y}_i|\hat{\theta}_i, \boldsymbol{\beta}_0; \mathcal{M}_0), \tag{3}$$

asymptotically follows a standard normal distribution when there is no person misfit. Special cases of $S^*(\hat{\theta}_i)$ include the popular $l_z^*(\hat{\theta}_i)$ statistic (Drasgow, Levine, & Williams, 1985; Snijders, 2001) and the extended caution index $\zeta_2^*(\hat{\theta}_i)$ (Sinharay, 2016a; Tatsuoka, 1984). Note that $S^*(\hat{\theta}_i)$ depends on the true IRFs, and is evaluated at the estimate $\hat{\theta}_i$.

As an alternative to a standard normal null distribution, a person-specific resampling-based null distribution may be used to obtain the p-value for $S^*(\hat{\theta}_i)$ (de la Torre & Deng, 2008; Sinharay, 2016b). This resampling procedure, which simulates many "observed" response patterns under the model, depends on the true IRFs as well as the estimate $\hat{\theta}_i$.

The Effect of Model Misspecification

Let $\mathcal{M}_*$ denote a misspecified IRT model. Under misspecification, let the population item parameters be $\boldsymbol{\beta}_*$, which will not equal $\boldsymbol{\beta}_0$. Let the corresponding vector of ML estimates be $\check{\boldsymbol{\beta}}$, which will not equal $\hat{\boldsymbol{\beta}}$. Finally, under misspecification, $\check{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}_*$ (White, 1982). For example, if $\mathcal{M}_0$ is the 3PL model, then $\mathcal{M}_*$ could be the 2PL or 1PL model. In this case, the numbers of parameters in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_*$ will not even be equal. Still, under misspecification, $\check{\boldsymbol{\beta}}$ will converge to a stationary population counterpart $\boldsymbol{\beta}_*$ as sample size increases (e.g., Falk & Monroe, 2018).

Notwithstanding $\boldsymbol{y}_i$, under misspecification, all quantities from the previous section are modified. More specifically, the IRF is

$$\tilde{P}_j(\theta_i) = \Pr(Y_{ij} = 1|\theta_i, \boldsymbol{\beta}_*; \mathcal{M}_*), \tag{4}$$

and let $\tilde{\theta}_i$ be an ability estimate based on the response pattern $\boldsymbol{y}_i$ and item parameters $\boldsymbol{\beta}_*$ that correspond to $\mathcal{M}_*$. We refer to $\tilde{P}_j(\theta_i)$ as a misspecified IRF and to $\tilde{\theta}_i$ as a misspecified ability estimate. Finally, under misspecification, the Snijders (2001) statistic in Equation 3 becomes

$$\tilde{S}^*(\tilde{\theta}_i) = S^*(\boldsymbol{y}_i|\tilde{\theta}_i, \boldsymbol{\beta}_*; \mathcal{M}_*). \tag{5}$$

Note that predicting the behavior of $\tilde{S}^*(\tilde{\theta}_i)$, in relation to $S^*(\hat{\theta}_i)$, is challenging, because $\tilde{S}^*(\tilde{\theta}_i)$ depends on misspecified IRFs, and is evaluated at a misspecified ability estimate.

In addition, the resampling-based null distribution approach is likewise affected by misspecification. In this case, the "observed data" will be simulated using misspecified IRFs, as well as the misspecified estimate $\tilde{\theta}_i$.

Simulation Study

A simulation study was conducted to examine the effect of item model misspecification on the Type I error rates of S^*. Previous simulation studies on S^* have

Table 1
Generating Parameters for Simulation Study

Item	1	2	3	4	5	6	7	8	9	10
a	.600	.913	1.300	1.285	.574	1.404	.796	1.162	1.033	1.421
b	−1.511	−1.354	−.619	−.008	.152	.219	.336	.866	.866	1.055
g	.000	.200	.060	.251	.183	.273	.130	.210	.176	.117

Note. Values reproduced from de la Torre and Deng (2008). For the 2PL and 1PL models, all g parameters are set to 0. Also, for the 1PL model, all a parameters are set to 1.

considered several factors, including the ability estimator, version of S^* (e.g., l_z^*), and null distribution. These factors were likewise included in the current study to examine whether some form of S^* might be robust to the studied misspecifications.

Design

The generating model (GM) was the 1PL, 2PL, or 3PL. To study S^* under correct model specification, the analysis model (AM) was specified to match the GM. To study the effects of misspecification, the following GM–AM combinations were studied: 2PL-1PL, 3PL-1PL, and 3PL-2PL.[2] To represent a moderate test length, 40 items were used.

Two sets of true data-generating item parameters were used. The first set was based on empirical item parameters from a mathematics test (de la Torre & Patz, 2005) previously studied in the person-fit literature (de la Torre & Deng, 2008; Sinharay, 2016b). The parameters for the 10 mathematics items, reproduced in Table 1, were repeated four times. The second set of item parameters was simulated based on Rupp (2013), who provided values commonly used in the person-fit simulation literature and that are arguably representative of large-scale educational assessments. Item discrimination parameters were randomly drawn from a Uniform (.75, 2) distribution. Difficulty parameters were drawn from a $N(0, 1)$ distribution truncated to the interval $[-2, 2]$. Finally, guessing parameters were drawn from a Uniform (0, .25) distribution. The sets of empirical and simulation-based generating parameters are referred to as $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$, respectively. Figure 1 shows the test information functions corresponding to $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$. Though the functions have similar shapes, $\boldsymbol{\beta}_0^S$ yields greater test information, and a greater marginal reliability (.85, compared to .79 for $\boldsymbol{\beta}_0^E$). For the 2PL and 1PL models, the relevant parameters were used, with $a = 1$ for all items for the 1PL model.

In total, there were six generating conditions, with two factors—GM (1PL, 2PL, or 3PL) and generating parameters ($\boldsymbol{\beta}_0^E$ or $\boldsymbol{\beta}_0^S$)—fully crossed. As described in the previous section, computation of S^* depends on the item parameters, and in many PFS studies the true item parameters, $\boldsymbol{\beta}_0$, are assumed known. However, in the current study, the population item parameters under misspecification, $\boldsymbol{\beta}_*$, are unknown. Thus, to make the comparison fair, item parameters were estimated under both the true and misspecified models. A single large calibration sample of $N = 5,000$, with $\theta \sim N(0,\ 1)$, was used so that the estimation uncertainty was negligible (i.e., $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0$ and

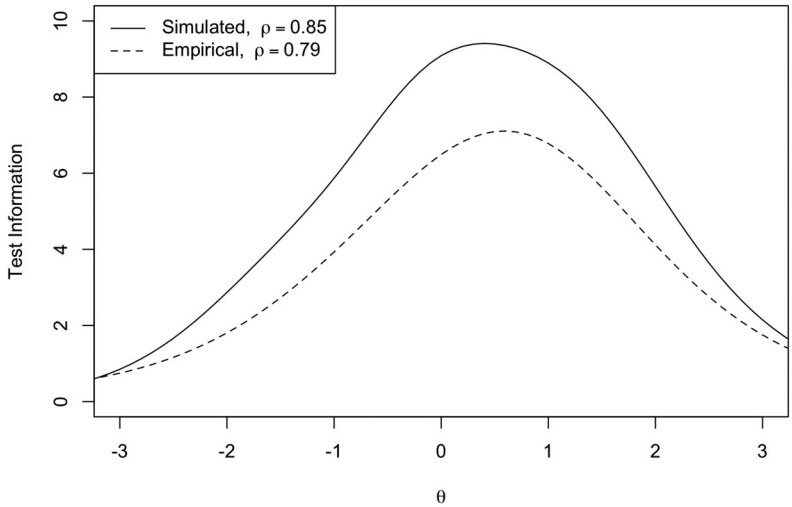

Figure 1. Test information functions for the 40 item tests.
Note. In the legend, ρ denotes the marginal reliability.

$\tilde{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_*$). Previous research suggests that the PFS results should not be meaningfully affected by this decision of whether to use true or estimated item parameters (Glas & Dagohoy, 2007; Molenaar & Hoijtink, 1990).

The PFSs were studied in aggregate and conditional on θ for each of the six generating conditions. To study the PFS in aggregate, a simulated data set of 5,000 examinees—independent of the calibration sample—was used, with $\theta \sim N(0,\ 1)$. To study the PFS conditional on θ, a simulated data set of 5,000 examinees was used for each of nine fixed values of θ, ranging from -2 to 2 in increments of .5. This number of examinees per fixed value of θ is equal to that used in de la Torre and Deng (2008).

For every response pattern, under each analysis model, eight PFSs were calculated. These differed based on the ability estimator, version of S^*, and null distribution. For the latent ability estimate, either the MAP or WLE was used. The S^* PFS has been studied most often using the MAP, WLE, and ML estimates (e.g., Magis, Raich, & Beland, 2012). However, ML can produce infinite θ estimates, but otherwise tends to correlate highly with the WLE estimate. Hence, ML was excluded from the study. Furthermore, the combination of MAP and WLE represents both Bayesian and likelihood frameworks. For the version of S^*, either l_z^* or ζ_2^* was used. These are among the most popular and best-performing variants of S^* (Sinharay, 2016a). To obtain the p-value, either the standard normal asymptotic null distribution (Snijders, 2001) or a resampling-based null distribution (de la Torre & Deng, 2008; Sinharay, 2016b) was used.[3]

Collected Statistics

Several model fit statistics were calculated for each condition. To assess the overall fit of the model, M_2 (Maydeu-Olivares & Joe, 2006) and the corresponding RMSEA (Maydeu-Olivares & Joe, 2014) were collected. And, as in Meijer and

Table 2
Model Fit Statistics for Simulation Study

Parameters	GM	AM	M_2 (df)	p-Value	RMSEA	$S-X^2$	X^2_{LD}
			Correctly specified model				
$\boldsymbol{\beta}_0^E$	1PL	1PL	752.36 (779)	.74	.000	.050	.051
	2PL	2PL	709.60 (740)	.78	.000	.025	.017
	3PL	3PL	677.17 (700)	.73	.000	.050	.024
$\boldsymbol{\beta}_0^S$	1PL	1PL	849.84 (779)	.04	.000	.000	.051
	2PL	2PL	735.20 (740)	.54	.000	.000	.016
	3PL	3PL	732.57 (700)	.19	.000	.025	.037
			Misspecified model				
$\boldsymbol{\beta}_0^E$	2PL	1PL	3060.23 (779)	<.01	.024	.800	.652
	3PL	1PL	1723.12 (779)	<.01	.016	.750	.385
	3PL	2PL	768.20 (740)	.23	.003	.200	.041
$\boldsymbol{\beta}_0^S$	2PL	1PL	2619.01 (779)	<.01	.022	.625	.582
	3PL	1PL	2776.51 (779)	<.01	.023	.650	.527
	3PL	2PL	852.70 (740)	<.01	.006	.125	.071

Note. GM = generating model; AM = analysis model. Values in $S-X^2$ and X^2_{LD} columns are rejection rates (out of 40 and 820 statistics) at $\alpha = .05$.

Tendeiro (2012), $S-X^2$ (Orlando & Thissen, 2000) was used to examine item fit, and X^2_{LD} (Chen & Thissen, 1997) was used to examine pairwise local dependence.

For evaluation of the PFSs, the Type I error rates at significance levels of .01 and .05 were collected. Item parameter calibration was performed using flexMIRT (Cai, 2017). When the AM was the 3PL, a $N(-1.09,\ 0.5)$ prior distribution was placed on the logit of the g parameter to stabilize estimation (e.g., Orlando & Thissen, 2000). The model fit statistics (e.g., M_2) were also obtained using flexMIRT. All other statistics were calculated using R (R Core Team, 2017).

Results

No estimation problems were encountered during the simulation study. Regarding the PFSs, the results are largely similar for the two ability estimators, MAP and WLE. To save space, only the results for the WLE estimator are presented.

Item Model Correctly Specified

The first rows of results in Table 2 summarize the model fit statistics when the item model is correctly specified, for both $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$. The overall goodness-of-fit statistic, M_2, and its corresponding root mean square error of approximation (RMSEA), indicate good fit for all combinations of item models and item parameters. Though guidelines for interpreting RMSEA for IRT models are still being refined (Maydeu-Olivares & Joe, 2014), all RMSEA values are 0, the minimum possible value. Regarding the item fit statistics, the rejection rates for $S-X^2$ are close to the nominal level ($\boldsymbol{\beta}_0^E$) or slightly conservative ($\boldsymbol{\beta}_0^S$). We note these rates should be

Table 3
Aggregate Type I Error Rates for Simulation Study

			l_z^*				ζ_2^*			
			Asymptotic		Resampling		Asymptotic		Resampling	
Parameters	GM	AM	.05	.01	.05	.01	.05	.01	.05	.01
Correctly specified model										
$\boldsymbol{\beta}_0^E$	1PL	1PL	.052	.013	.044	.009	.051	.012	.047	.009
	2PL	2PL	.053	.014	.045	.008	.049	.011	.044	.008
	3PL	3PL	.052	.013	.044	.008	.050	.012	.044	.008
$\boldsymbol{\beta}_0^S$	1PL	1PL	.060	.018	.055	.012	.059	.017	.056	.011
	2PL	2PL	.058	.014	.054	.010	.052	.012	.052	.010
	3PL	3PL	.052	.018	.051	.011	.047	.014	.052	.012
Misspecified model										
$\boldsymbol{\beta}_0^E$	2PL	1PL	.064	.018	.056	.011	.063	.016	.058	.011
	3PL	1PL	.080	.023	.071	.017	.067	.017	.062	.013
	3PL	2PL	.066	.017	.058	.012	.061	.015	.055	.011
$\boldsymbol{\beta}_0^S$	2PL	1PL	.057	.015	.052	.010	.059	.016	.059	.013
	3PL	1PL	.065	.022	.063	.017	.063	.021	.065	.019
	3PL	2PL	.059	.020	.057	.015	.050	.015	.055	.013

Note. GM = generating model; AM = analysis model. Values are underlined if they are statistically greater than the nominal rate.

interpreted along with the number of items, because fewer items will lead to more variable rates under the null hypothesis. Regarding the pairwise fit statistics, the rejection rates for X_{LD}^2 are either close to the nominal level (1PL) or slightly conservative (2PL and 3PL). These results are consistent with prior research on X_{LD}^2 (e.g., Hansen, Cai, Monroe, & Li, 2016). In summary, the model fit statistics all suggest acceptable fit, which is unsurprising given the correct specification.

Table 3 is organized similar to Table 2, and the first rows present the aggregate Type I error rates for the correctly specified models, for both $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$. For the statistics based on the asymptotic null distribution, the Type I error rates for both l_z^* and ζ_2^* are inflated at $\alpha = .01$, which is consistent with previous research (e.g., Snijders, 2001) and expected, given the moderate length of the test. In comparison, the resampling null distribution leads to better Type I error control, which again is consistent with previous research (e.g., de la Torre & Deng, 2008). Finally, in comparing the rates of l_z^* and ζ_2^*, the latter statistic exhibits less Type I error inflation, which was also found in Sinharay (2016a). These patterns of results hold for both $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$, and overall, the statistics are well calibrated.

Figures 2 and 3 show conditional Type I error rates for the correctly specified models, for $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$, respectively. The gray lines represent 95% confidence intervals for the rates, based on a normal approximation to the binomial distribution.[4] For either $\boldsymbol{\beta}_0^E$ or $\boldsymbol{\beta}_0^S$, and any of the statistics (e.g., l_z^* and the asymptotic null distribution), there

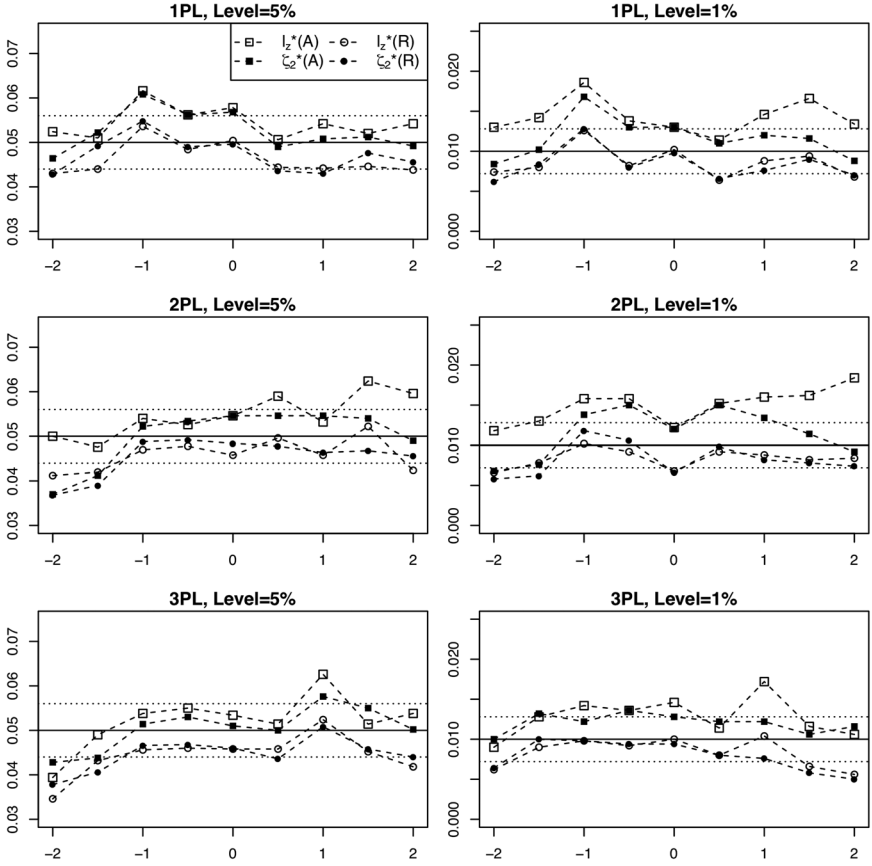

Figure 2. Type I error rates for correctly specified models for $\boldsymbol{\beta}_0^E$ item parameters.
Note. In the legend, "(A)" denotes the PFS uses the asymptotic null distribution and "(R)" denotes the PFS uses the resampling-based null distribution. Black dotted lines demarcate 95% confidence intervals.

is clear variability in the rates across the ability range. However, the Type I error rates are never too different from the nominal rates. At $\alpha = .01$, the rates range from .005 to .020, and at $\alpha = .05$, the rates range from .03 to .07.

Item Model Misspecified

The effects of misspecification were studied by specifying a simpler AM than GM. It was predicted that among the studied conditions, the 3PL–1PL combination would lead to the worst PFS performance. Also, it was expected that conditional Type I error rates would be the worst calibrated for extreme latent trait values. This is because, when the AM is simpler than the GM, a large proportion of the misfit in the IRF often occurs at extreme values of θ (Orlando & Thissen, 2000; Sinharay, 2006). It was reasoned that the misspecification would therefore disproportionately affect the corresponding conditional PFS Type I error rates.

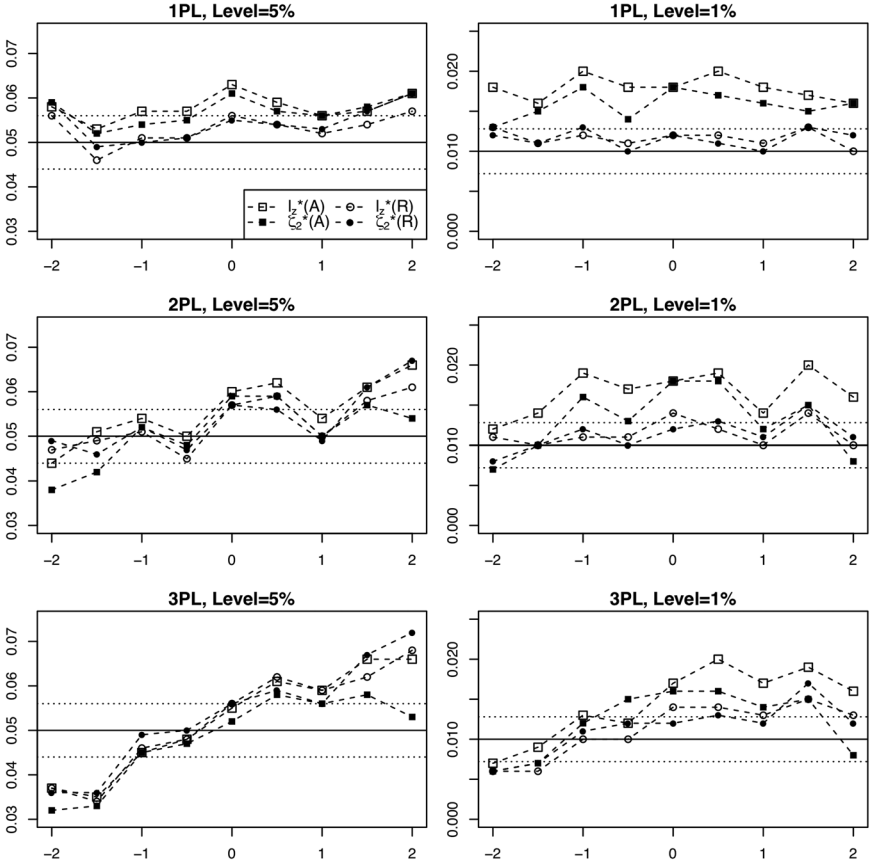

Figure 3. Type I error rates for correctly specified models for $\boldsymbol{\beta}_0^S$ item parameters.
Note. In the legend, "(A)" denotes the PFS uses the asymptotic null distribution and "(R)" denotes the PFS uses the resampling-based null distribution. Black dotted lines demarcate 95% confidence intervals.

Table 2 also presents model fit statistics for the misspecified models, for both sets of true item parameters, $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$. The results, however, are quite similar for $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$. For the 2PL–1PL and 3PL–1PL combinations, the statistics strongly suggest the 1PL does not provide good fit. For these conditions, the M_2 statistic is significant, and the RMSEA values are all .02. In the structural equation modeling literature, an RMSEA of .02 is often interpreted as indicating "close" fit (Browne & Cudeck, 1993), but more stringent guidelines may be necessary for categorical data models (Monroe & Cai, 2015). More clearly, the large rejection rates for $S - X^2$ (ranging from .63 to .80) and X^2_{LD} (ranging from .39 to .65) indicate the 1PL provides poor fit.

In contrast, for the 3PL–2PL combination, the statistics do not indicate the misfit as clearly. For both $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$, M_2 is not significant and the RMSEA is less than .01. Additionally, the X^2_{LD} rejection rates are near .05. Both of these results

are similar to the corresponding results for the correctly specified models, and support previous findings that when data are generated from the 3PL model the 2PL model tends to fit the data reasonably well (Orlando & Thissen, 2000; Sinharay, 2006). On the other hand, the rejection rates for $S - X^2$ are slightly higher for the 3PL–2PL combination than those for the correctly specified models, with rates of .2 for $\boldsymbol{\beta}_0^E$, and .13 for $\boldsymbol{\beta}_0^S$, which correspond to eight and five flagged items, respectively.

Table 3 also presents aggregate Type I error rates for the misspecified models. In comparison to the rates for the correctly specified models, the corresponding rates under misspecification are all slightly inflated. The highest rates at the .01 and .05 significance levels are .023 and .08, respectively, which both occur for $\boldsymbol{\beta}_0^E$, the 3PL–1PL combination, and l_z^* with the asymptotic null distribution. However, for several entries, the rates are not significantly greater than the nominal levels. Thus, although the aggregate rates are mostly inflated under misspecification, the degree of inflation is generally modest.

Figures 4 and 5 show the Type I error rates for fixed θ under misspecification, for $\boldsymbol{\beta}_0^E$ and $\boldsymbol{\beta}_0^S$, respectively. The figures share several features. First, in comparison to Figures 2 and 3, the Type I error rates under misspecification show clearer patterns. For example, in Figures 4 and 5, for the 2PL–1PL combination, the rates are conservative for low θ values, and clearly increase with θ. Second, relatedly, the various PFSs are similarly affected by the misspecification. In other words, for each of the plots, the general pattern does not depend on the version of S^* or the null distribution. Third, unlike the aggregate rates under misspecification (see Table 3), the rates at fixed values of θ may differ greatly from the nominal rates. As an extreme example, in Figure 5, for the 3PL–1PL combination at $\alpha = .05$, the rates range from approximately .02 to .27. Fourth, overall, the magnitude of the difference of the Type I error rates from the nominal levels depends on the GM and AM. The differences are greatest for the 3PL–1PL combination, and smallest for the 3PL–2PL combination. Fifth, the Type I error rates are most likely to be inflated for extreme values of θ. In contrast, for central values of θ (i.e., $-0.5 \leq \theta \leq 0.5$), the Type I error rates were fairly accurate.

There are also several differences between Figures 4 and 5. For the 2PL–1PL and 3PL–1PL combinations, for low θ values, the rates for $\boldsymbol{\beta}_0^E$ are lower than those for $\boldsymbol{\beta}_0^S$. In contrast, for the 3PL–1PL combination, for high θ values, the rates for $\boldsymbol{\beta}_0^E$ are higher than those for $\boldsymbol{\beta}_0^S$. A final difference is that the variability in the rates, across combinations of S^* and null distribution, is somewhat greater for $\boldsymbol{\beta}_0^E$ than for $\boldsymbol{\beta}_0^S$. That is, for $\boldsymbol{\beta}_0^E$, the rates are relatively sensitive to the version of S^* and null distribution.

In summary, the results under misspecification were consistent with the expectations that the PFSs would perform worst for the 3PL–1PL combination, and that the Type I error rates would be most distorted for extreme values of θ. However, the results provided numerous additional details on how misspecification can affect S^*. Notably, results showed that substantial inflation in conditional rejection rates could result from misspecification, despite relatively well-controlled aggregate rejection rates, and only minor indications of misfit (for the 3PL–2PL combination). Additionally, the simulation study confirmed that it is challenging to predict the effects of misspecification on S^*.

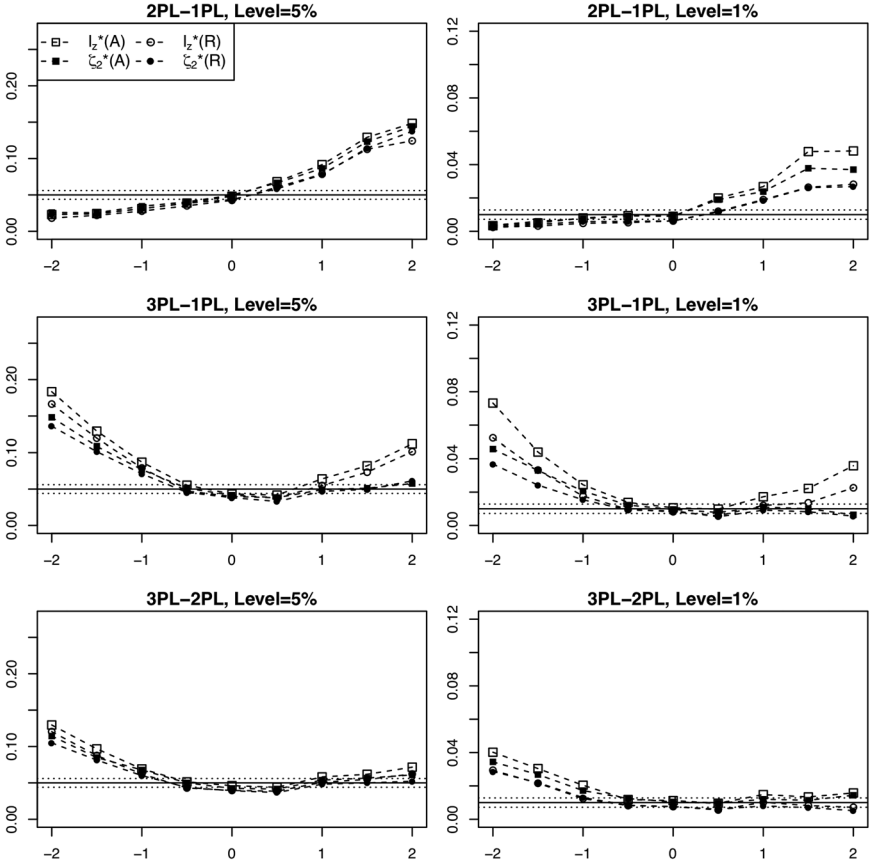

Figure 4. Type I error rates for misspecified models for $\boldsymbol{\beta}_0^E$ item parameters.
Note. Plot titles provide generating model, followed by analysis model. In the legend, "(A)" denotes the PFS uses the asymptotic null distribution and "(R)" denotes the PFS uses the resampling-based null distribution. Black dotted lines demarcate 95% confidence intervals.

Empirical Examples

Two data sets were analyzed to illustrate the effects of model specification on PFSs. The first data set is from a state math assessment, and the second is from a licensure exam. For both data sets, the 1PL, 2PL, and 3PL models were each used as the AM. In addition to the model fit statistics calculated in the simulation study, the Bayesian information criterion (BIC) was also calculated to inform model selection. For the PFS, l_z^*, the WLE, and the asymptotic null distribution were used.

State Math Assessment

The data are item responses from $N = 5,000$ fifth-grade students to 44 items, scored dichotomously. The assessment is not high-stakes for the students. The state

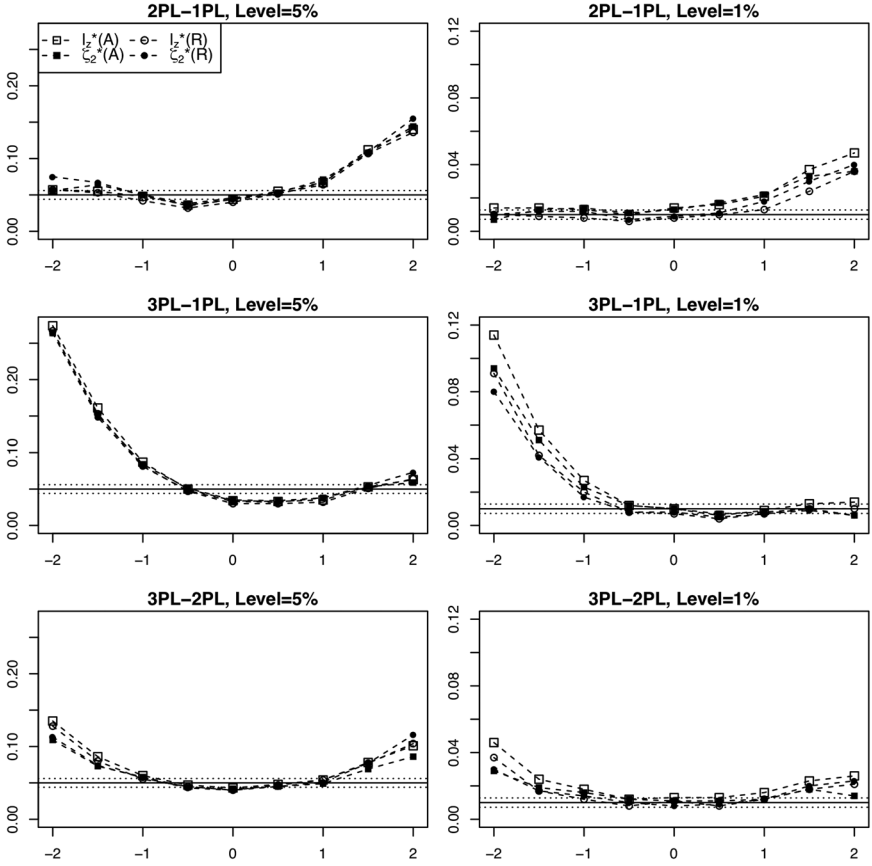

Figure 5. Type I error rates for misspecified models for $\boldsymbol{\beta}_0^S$ item parameters.
Note. Plot titles provide generating model, followed by analysis model. In the legend, "(A)" denotes the PFS uses the asymptotic null distribution and "(R)" denotes the PFS uses the resampling-based null distribution. Black dotted lines demarcate 95% confidence intervals.

is not identified for legal reasons and the operational item model is similarly not disclosed.

Table 4 reports the model fit statistics for the three AMs. The 3PL model provides the best absolute fit, as judged by M_2. But it also provides the best fit after accounting for model complexity, according to the BIC and RMSEA values. Moreover, the rejection rates at $\alpha = .05$ for the $S - X^2$ and X^2_{LD} statistics indicate the 3PL model fits the data substantially better than the 2PL or 1PL models. While these rates are somewhat inflated for the 3PL (.11 and .14 for $S - X^2$ and X^2_{LD}, respectively), they are much closer to the nominal level than the rates for the 2PL (.50 and .47) or 1PL models (.82 and .70).

Table 4 also reports the aggregate rejection rates for l_z^* at the .01 and .05 significance levels. The rates for the 3PL model are closest to the nominal rates. If the 3PL

Table 4
Empirical Data Analysis Results

	Model Fit Statistics						l_z^*	
Model	BIC	M_2 (df)	p-Value	RMSEA	$S-X^2$	X^2_{LD}	.05	.01
			State Mathematics Assessment					
1PL	255216.8	6647.03 (945)	<.001	.035	.818	.607	.100	.042
2PL	253565.7	4933.55 (902)	<.001	.030	.500	.467	.087	.031
3PL	252544.3	3207.92 (858)	<.001	.023	.114	.142	.075	.025
			Licensure Test					
1PL	64382.8	1665.82 (779)	<.001	.026	.375	.207	.059	.015
2PL	64380.5	1292.13 (740)	<.001	.021	.150	.076	.047	.014
3PL	64670.3	1170.81 (700)	<.001	.020	.125	.079	.044	.015

Note. Values in $S-X^2$ and X^2_{LD} columns are rejection rates at $\alpha = .05$. Values in l_z^* columns are rejection rates. BIC = Bayesian information criterion; RMSEA = root mean square error of approximation.

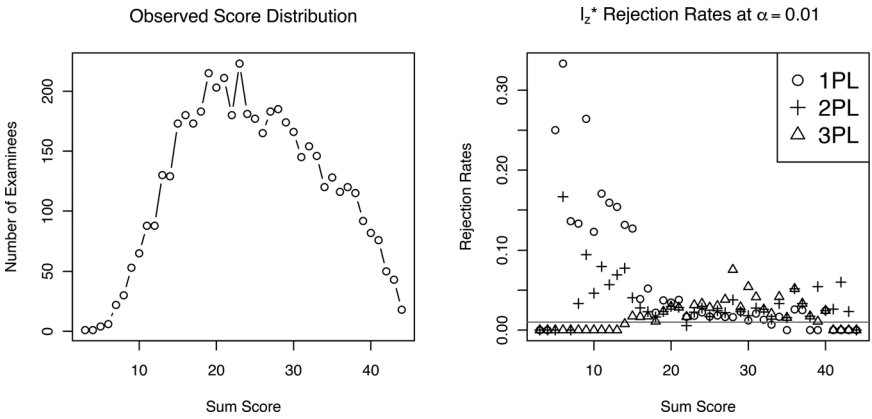

Figure 6. Examinee sample size and PFS rejection rates for state assessment.

is considered well fitting, and the 1PL and 2PL models are considered misspecified, then the elevated rates for these latter models can be attributed to model misspecification (Meijer & Tendeiro, 2012). However, it could be argued that the differences in the rates are not overly meaningful (e.g., the rates for the 3PL and 2PL at $\alpha = .01$ are .025 and .031, respectively).

Figure 6 displays two statistics conditional on the observed sum score, denoted as T: the sample size (left) and rejection rate at $\alpha = .01$ (right). The plot of rejection rates may be considered an analog to Figures 2–5, with T in place of the latent ability θ. The sum score T is used as a convenient grouping variable for rejection rates, because unlike $\hat{\theta}$, T does not depend on the item model. Thus, for empirical analyses, conditioning on T facilitates comparisons across models. Notably, for each model, the rates are not constant across the T. For students with low scores (say,

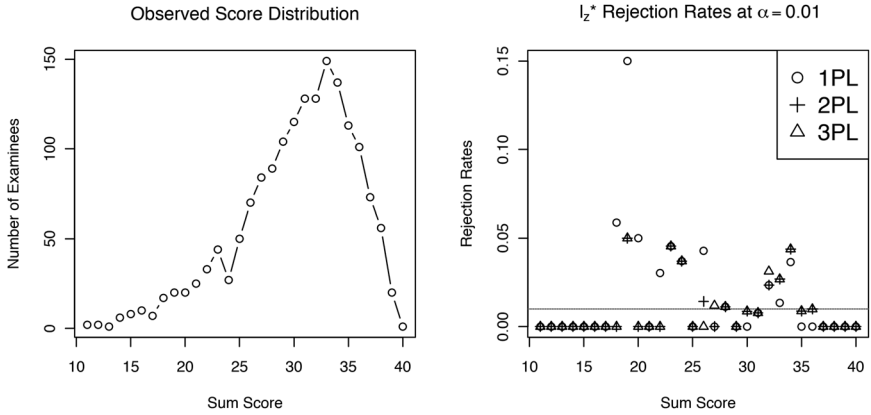

Figure 7. Examinee sample size and PFS rejection rates for licensure exam.

$T \leq 15$), the rejection rates using the 1PL and 2PL models are much greater than the rates using the 3PL model. On the other hand, for students with higher scores (say, $T > 15$), the rejection rates using the different models are more similar. Note that the aggregate rates in Table 4 fail to reveal this variability. Again, if the 1PL and 2PL models are considered misspecified, then the inflated rates for low T can be attributed to the misspecification. From another perspective, for this data set, l_z^* is sensitive to the AM, at least for low-ability examinees.

Licensure Exam

The data are from a licensure exam form with $N = 1,644$ examinees and 170 dichotomously scored operational items. This data set has been analyzed previously in the PFS literature, notably in several chapters in Cizek & Wollack (2017). Operationally, a 1PL model is used for the exam. Due to the focus of the current research, a subset of the items was analyzed, as the sample size was too small to provide accurate item parameter estimation for the full set of 170 items (e.g., De Ayala, 2013). To align the example with the simulation study, 40 items were randomly chosen from among the subset of items with corrected item-total correlations greater than .2.[5]

Table 4 also reports the model fit statistics for the licensure exam. Interestingly, the BIC values for the 1PL and 2PL models are approximately equal, and both lower than the 3PL BIC value. However, the 3PL model is preferred by the RMSEA. Thus, determining which model fits best using these overall indices is challenging. On the other hand, for the 2PL and 3PL models, the $S - X^2$ and X^2_{LD} rejection rates are very similar (approximately .15 and .08, respectively). In contrast, the corresponding rates for the 1PL model are substantially greater (.38 and .21, respectively). Based on these results, it appears that the 2PL and 3PL models provide comparable fit, and both are to be preferred over the 1PL.

The aggregate l_z^* rejection rates, reported in Table 4, are similar across the models. The conditional rates for $\alpha = .01$ are displayed in Figure 7. Unlike with the math assessment analysis in the previous example, the 2PL and 3PL models yield very similar rates across T. In contrast, the rates for the 1PL tend to be greater (e.g.,

$T = 19$). Based on the comparable model fit statistics and conditional rejection rates, it appears that l_z^* is not overly sensitive to the choice between the 2PL and 3PL models. However, based on these criteria, the 1PL should not be used to compute l_z^* for this analysis.

Conclusion

In this research, the effects of misspecification on the performance of PFSs in the absence of aberrant behavior was explored. A simulation study was conducted to investigate the relationships between model specification, model fit evaluation, and aggregate and conditional rejection rates. The simulation study also varied the ability estimator, version of S^*, and the reference null distribution to see whether any combination of these factors would prove robust to the studied misspecifications.

The simulation study demonstrated that aggregate Type I error rates under misspecification were inflated when compared to the rates under correct model specification. However, this inflation was rather modest (up to .08 at $\alpha = .05$). On the other hand, the inflation of the conditional rates under misspecification was relatively dramatic at times (up to .27 at $\alpha = .05$). The impact of misspecification is harder to detect with the aggregate rates because the greatest inflation tended to occur for extreme values of θ, which occur relatively infrequently. The simulation study also showed that none of the studied versions of S^* were robust to the misspecification. With that said, consistent with previous findings, use of a resampling-based null distribution (de la Torre & Deng, 2008; Sinharay, 2016b) yielded better Type I control than the asymptotic null distribution for S^*.

The 2PL–1PL and 3PL–1PL combinations exhibited poor model fit, and also had the greatest degree of inflation. This relationship is reassuring in the sense that, for these combinations, the poor performance of the PFSs may be predicted by poor model fit. For the 3PL–2PL combination, there was less evidence of poor model fit. Nevertheless, this combination resulted in inflated conditional rates for some extreme values of θ, especially those at the low end under the most realistic conditions in our simulation study. This relationship suggests that an adequately fitting model, as judged by some fit statistics, may still lead to poor PFS performance.

These results are reconcilable with Meijer and Tendeiro (2014), who used nonparametric response functions and PFSs with high-stakes educational testing data. These authors noticed that respondents consistently flagged using the studied PFSs tended to have relatively low scores, and the authors concluded that this relationship might be due to extensive respondent guessing. Our results are complementary in that we found similar patterns, but used parametric response functions and PFSs, with both real and simulated data. These results have implications for equity to the extent that those with low ability are potentially singled out more often as problematic responders.

Based on our findings from the simulation study and empirical analyses, we encourage practitioners to consider specifying alternative, relatively complex, item models for purposes of producing PFSs. Though this may be inconvenient in some operational settings, it is arguably preferable to relying on PFSs of questionable quality. Further, we encourage practitioners to examine conditional PFS rejection rates

across alternative models. For instance, we used a graphical approach for studying conditional rejection rates of PFSs with empirical data. Figures 6 and 7 show how rejection rates can be compared across item models, conditioning on T instead of θ. In practice, such plots could be constructed with more complex models for all items, with mixed format tests, or with alternative models that might fit the data better than the 3PL model (e.g., Culpepper, 2017; Lee & Bolt, 2018). Similarly, whether individual respondents are consistently flagged across alternative models can also be examined. Unfortunately, whether differences in rejection rates or decisions for individuals are substantial enough to warrant caution is somewhat subjective and may depend on the particular testing situation. With that said, this recommended strategy examines the practical impact of misspecification of IRT models (Sinharay & Haberman, 2014) on PFSs.

We also note that regardless of how a PFS is computed, it alone cannot provide conclusive proof of aberrant test-taking behavior. Instead, multiple sources of information (e.g., interviews, seating charts, respondent history, notes in testing booklets, and so on) should be collected and considered. Studies exemplifying such efforts, such as Meijer, Egberink, Emons, and Sijtsma (2008), are important for practitioners, and more are needed.

With regard to this study, there are numerous directions for future research. First, other simulation conditions can be considered, such as different test lengths or sample sizes. In addition, PFSs besides S^* may behave differently under misspecification, and such alternative PFSs could be studied. Second, new PFSs may be developed that are less sensitive to misspecification. Although nonparametric PFSs are a promising alternative, further research, along the lines of Emons (2008), is needed to develop approaches that lead to well-calibrated statistics conditional on T or θ, in particular for extreme scores. Third, other forms of model misspecification may be considered (e.g., multidimensionality). Fourth, other methods of model fit evaluation may prove useful in the current context. Fifth, the effects of misspecification on the power of PFSs should be studied. We expect that under misspecification, differentiating between false and true positives will be more difficult, especially at extreme levels of the latent trait. Finally, though we have made suggestions for practitioners, more research is needed on how to use PFSs in practical situations.

Acknowledgments

We thank Dr. Sandip Sinharay for his feedback on an early version of this manuscript. We would also like to thank Dr. James Wollack for generously providing an empirical data set analyzed in this research.

Notes

[1]Both Snijders (2001) and de la Torre and Deng (2008) studied the l_z^* statistic.

[2]GM–AM combinations with a more complex AM than GM were not studied because it was expected that such combinations would yield similar results to the correctly specified combinations, due to the large calibration sample size. With more parameters in the AM than in the GM, the additional parameter estimates should

converge to their population counterparts (e.g., for the 1PL–2PL combination, the item-specific slope estimates should converge to the true common slope value).

[3]For the resampling scheme, for each combination of response pattern and analysis model, 1,000 values of θ were simulated from a normal distribution with mean and *SD* equal to the examinee's ability estimate and *SE*, respectively. In Sinharay (2016b), this resampling scheme is labeled the "Monte Carlo" approach. Then, each of these θ values was used to simulate a response pattern based on the AM and corresponding item parameter estimates. Next, each response pattern was used to calculate an ability estimate and PFS. The 1,000 values of the PFS constitute the examinee-specific resampling-based distribution. The *p*-value is computed as the proportion of resampling-based PFS values more extreme than the examinee's PFS value. For further details, see Sinharay (2016b, p. 68).

[4]For instance, using the binomial approximation, with 5,000 replications the 95% confidence interval is (.008, .013) for $\alpha = .01$, and is (.044, .056) for $\alpha = .05$.

[5]Though these data have been analyzed previously in the PFS literature, our decision to use a subset of the items makes comparisons with these previous analyses challenging.

References

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136–136). Newbury Park, CA: Sage.

Cai, L. (2017). *flexMIRT® version 3.5: Flexible multilevel multidimensional item analysis and test scoring. [Computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Conijn, J. M., Emons, W. H., & Sijtsma, K. (2014). Statistic l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*, 122–136.

Crisan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*, 439–455.

Culpepper, S. A. (2017). The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics*, *42*, 706–725.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford Press.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159–177.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, *6*, 297–308.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247.

Falk, C. F., & Monroe, S. (2018). On Lagrange multiplier tests in multidimensional item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement*, *78*, 653–678.

Glas, C. A., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Lee, S., & Bolt, D. M. (2018). An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects. *Journal of Educational Measurement*, *55*, 90–111.

Magis, D., Raich, G., & Beland, S. (2012). A didactic presentation of Snijers's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57–81.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305–328.

Meijer, R. R., Egberink, I. J., Emons, W. H., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, *90*(3), 227–238.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the l_z and l_z^* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*, 758–766.

Meijer, R. R., & Tendeiro, J. N. (2014). *The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us* (LSAC Research Report Series). Newtown, PA: Law School Admission Council.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75–106.

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*, 569–583.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

R Core Team, (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3–8.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429–449.

Sinharay, S. (2016a). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement*, *40*, 418–433.

Sinharay, S. (2016b). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement*, *53*, 63–85.

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.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95–110.

Van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327–345.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*, 484.

Authors

SEONG EUN HONG is a PhD student in Research, Educational Measurement and Psychometrics, University of Massachusetts Amherst, 813 N. Pleasant St., Amherst, MA 01003; shong@umass.edu. Her primary research interests include item response theory and structural equation modeling.

SCOTT MONROE is Assistant Professor in Research, Educational Measurement and Psychometrics, University of Massachusetts Amherst, 813 N. Pleasant St., Amherst, MA 01003; smonroe@educ.umass.edu. His primary research interests include item response theory and structural equation modeling.

CARL F. FALK is Assistant Professor of Quantitative Methods and Modeling at McGill University, Montreal, QC H3A 1G1, Canada; carl.falk@mcgill.ca. His research focuses primarily on the development, programming, and testing of latent variable models.