PSYCHOMETRIKA—VOL. 85, NO. 2, 502–525
JUNE 2020
https://doi.org/10.1007/s11336-020-09712-7

POWER DIVERGENCE FAMILY OF STATISTICS FOR PERSON PARAMETERS IN IRT MODELS

XIANG LIU

COLUMBIA UNIVERSITY

EDUCATIONAL TESTING SERVICE

JAMES YANG, HUI SOO CHAE AND GARY NATRIELLO

COLUMBIA UNIVERSITY

We generalize the power divergence (PD) family of statistics to the two-parameter logistic IRT model for the purpose of constructing hypothesis tests and confidence intervals of the person parameter. The well-known score test statistic is a special case of the proposed PD family. We also prove the proposed PD statistics are asymptotically equivalent and converge in distribution to χ_1^2. In addition, a moment matching method is introduced to compare statistics and choose the optimal one within the PD family. Simulation results suggest that the coverage rate of the associated confidence interval is well controlled even under small sample sizes for some PD statistics. Compared to some other approaches, the associated confidence intervals exhibit smaller lengths while maintaining adequate coverage rates. The utilities of the proposed method are demonstrated by analyzing a real data set.

Key words: power divergence family, IRT, sufficient statistics, asymptotic distribution, interval estimation.

1. Introduction

Item response theory (IRT; Hambleton & Swaminathan, 1985) assumes that a set of categorical manifest variables measures a single latent trait. Given a response pattern, an essential problem is the estimation of the latent trait (person parameter). In practice, the latent trait is often estimated in two stages. The item parameters are first estimated from a calibration sample usually by marginal maximum likelihood (Baker & Kim, 2004). Under large calibration sample sizes, the sampling error of the item parameters is negligible (Liu & Yang, 2018). In other words, the item parameters can be estimated accurately which leads to the plug-in method for estimating the person parameter. In the second stage, treating the item parameters as fixed, the person parameter can be estimated with the maximum likelihood estimator (MLE).

Under finite sample sizes, the person parameter is estimated with error. The MLE is approximately normally distributed with the asymptotic variance given by the inverse of the Fisher information if the number of items is sufficiently large (Liu & Yang, 2004). The formal discussion of this well-known first-order variance of the MLE of the person parameter can be dated back at least to Lord (1983). Following the asymptotic normality, one can construct a Wald statistic and perform hypothesis testing of the person parameter. Moreover, the associated confidence intervals can be obtained by inverting the hypothesis tests (Casella & Berger, 2001). However, the asymptotic normality is often found to be unattainable under more realistic conditions where the number of items is moderate (e.g. Lord, 1983; Warm, 1989). Robust inference procedures

Xiang Liu and James Yang have contributed equally.

Correspondence should be made to Xiang Liu, Educational Testing Service, 03-T, 660 Rosedale Road, Princeton, NJ08541, USA. Email: xliu003@ets.org

that do not depend on the asymptotic normality have been discussed in more recent literature. For example, observing the raw sum score is sufficient for estimating the person parameter in the Rasch model, Klauer (1991) based inference on the exact distribution of the raw sum score. Similarly, Liu, Han, and Johnson (2018) extended the approach to the two-parameter logistic (2PL) model. Instead of the raw sum score, they calculated the exact distribution of the weighted sum score. In addition, a branch and bound algorithm is introduced to alleviate the computational difficulties associated with the approach. Similar exact distribution methods have been explored by Doebler, Doebler, and Holling (2012). They introduced a likelihood ratio-type statistic and a Wald-type statistic whose distributions were calculated based on the exact distribution of the finite responses (see Liu et al., 2018 for a detailed review). On the other hand, Biehler, Holling, and Doebler (2014) considered an higher-order approximation of the distribution of the weighted sum score which was computationally less demanding. More generally, Ogasawara (2012) derived a higher-order asymptotic expansion for the MLE of the ability parameter up to the fourth cumulant for the three-parameter logistic (3PL) model. Subsequent development extended the similar expansion to the Bayes and pseudo Bayes estimators of the ability parameter (Ogasawara, 2013). These higher-order methods tend to result in faster convergence to the true distribution as the number of items increases. As a result, the confidence intervals based on these methods usually provide coverage rates very close to the nominal level even under small to moderate item lengths.

Many of the aforementioned methods of constructing hypothesis tests and the associated confidence intervals of the person parameter can be related to the Wald test of the MLE and its exact distribution extensions. In statistical literature, other types of statistics have been widely discussed. Examples include the log-likelihood ratio statistic (G^2), Pearson's Chi-square statistic (χ^2), the Freeman–Tukey statistic (T^2), the Neyman modified Chi-square statistic (NM^2) and the modified log-likelihood ratio statistic (MG^2). More importantly, these statistics are special cases of the power divergence (PD) family of statistics introduced by Cressie and Read (1984) for multinomial models. The PD family of statistics is a subfamily of the more general ϕ-divergences (Morimoto, 1963). However, discussion of this topic in behavioral sciences is relatively sparse (e.g. Basu, Shioya, & Park, 2011; Felipe, Miranda, & Pardo, 2015; Ogasawara, 2019).

In the current paper, we generalize the PD family of statistics to the 2PL model for the purpose of constructing hypothesis tests and confidence intervals of the person parameter. The PD family includes the score test statistic as an important special case. We also prove the proposed PD statistics are asymptotically equivalent to each other and converge in distribution to χ^2_1. In addition, a moment matching method is introduced to compare statistics and choose the optimal one within the PD family such that the associated confidence intervals can provide adequate coverage even for smaller number of items. The finite sample properties of the resulting confidence intervals are compared with other alternative via simulations, and empirical data examples are presented in the end.

2. The PD Statistics for the IRT Models

2.1. 2PL, the Sufficient Statistic, and the Wald Test

Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_J)$ be item discrimination parameters and item difficulty parameters. Under the 2PL, the probability of a correct response conditional on some latent variable θ is given by

$$P(X_j = 1|\theta) = p_j(\theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}. \tag{1}$$

Given a response pattern $\boldsymbol{x}$, the likelihood function for θ is

$$L(\theta|\boldsymbol{x}) = \prod_{j=1}^{n} p_j^{x_j}(1-p_j)^{1-x_j} = \prod_{j=1}^{n} \left\{\frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]}\right\}^{x_j} \left\{\frac{1}{1+\exp[a_j(\theta-b_j)]}\right\}^{1-x_j}. \tag{2}$$

The likelihood function in (2) can be further factorized into its exponential family form,

$$L(\theta|\boldsymbol{x}) = \exp\left[\theta\sum_{j=1}^{J} a_j x_j\right]\left\{\exp\left[\sum_{j=1}^{J} a_j x_j b_j\right]\right\}^{-1}\left\{\prod_{j=1}^{J}\left\{1+\exp[a_j(\theta-b_j)]\right\}\right\}^{-1}. \tag{3}$$

It follows that $T(\boldsymbol{x}) = \sum_{j=1}^{J} a_j x_j$ is a sufficient statistic for θ. This result has been discussed before. For example, Harel (2014) provided a proof that the weighted sum score is a minimal sufficient statistic for θ under the more General Partial Credit Model (GPCM) through a direct Fisher–Neyman factorization. The uniformly most powerful (UMP) test and the associated computational algorithm suggested by Liu et al. (2018) also take advantage of the sufficiency of the weighted sum score for estimating θ.

Differentiate the log-likelihood function of θ,

$$U(\theta) = \frac{\partial \log L}{\partial \theta} = \sum_{j=1}^{J} a_j x_j - \sum_{j=1}^{J} a_j p_j(\theta). \tag{4}$$

The maximum likelihood estimator (MLE), $\hat{\theta}$, is then obtained by solving $U(\theta) = 0$ for θ. Unfortunately, no closed form solution for $\hat{\theta}$ is available. As a result, we have to find it numerically (Baker & Kim, 2004). In real world, we only observe a finite number of responses from a person. Hence, the latent trait θ is always estimated with error. Under the assumption that the number of items is sufficiently large and the true latent trait $\theta = \theta_0$ is away from the boundary of the parameter space, the MLE $\hat{\theta}$ is approximately normally distributed with a mean of θ_0 and some variance $var(\hat{\theta})$. In other words, the following Wald statistic

$$t_1 = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{var}(\hat{\theta})}} \tag{5}$$

has an asymptotic standard normal distribution. Equivalently, $t_1^2 \xrightarrow{d} \chi_1^2$. In practice, the variance of MLE is often approximated by the inverse of Fisher information evaluated at $\hat{\theta}$, i.e. $var(\hat{\theta}) \approx \left(-\frac{\partial^2 \log L}{\partial \theta^2}\right)^{-1}|_{\theta=\hat{\theta}}$. Comparing to some threshold value, t_1^2 can be used to test the linear hypothesis $H_0 : \theta_0 - \theta = 0$. Capitalizing on the sufficiency of $T(\boldsymbol{x}) = \sum_{j=1}^{J} a_j x_j$ for estimating θ, an alternative approach is to test a nonlinear hypothesis of the form $H_0 : c(\theta) = \sum_{j=1}^{J} a_j p_j(\theta_0) - \sum_{j=1}^{J} a_j p_j(\theta) = 0$. Notice that $\sum_{j=1}^{J} a_j p_j(\hat{\theta}) = \sum_{j=1}^{J} a_j x_j$. The Wald statistic for testing the nonlinear hypothesis is then

$$
\begin{aligned}
t_2^2 &= \frac{\left(\sum_{j=1}^{J} a_j x_j - \sum_{j=1}^{J} a_j p_j(\theta_0)\right)^2}{c'(\hat{\theta})^2 var(\hat{\theta})} \\
&= \frac{\left(\sum_{j=1}^{J} a_j x_j - \sum_{j=1}^{J} a_j p_j(\theta_0)\right)^2}{\sum_{j=1}^{J} a_j^2 p_j(\hat{\theta})[1 - p_j(\hat{\theta})]}.
\end{aligned} \tag{6}
$$

In general, given a response pattern and θ_0, t_1^2 and t_2^2 will not be exactly the same. Confidence interval of θ is most commonly constructed by inverting the Wald-test using t_1.

2.2. PD Statistics

The original PD family of statistics was introduced for multinomial models. To indicate how observed multinomial variables X_i for $i = 1, 2, \ldots, K$ categories differ from their expected values $E_i = n\pi_{0i}$, the PD statistic[1] is defined as

$$
2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{K} X_i \left\{ \left(\frac{X_i}{E_i} \right)^{\lambda} - 1 \right\}, \lambda \in \mathbb{R}. \tag{7}
$$

The problem we are considering here has some subtle but important differences to that of Cressie and Read (1984). Under IRT models, conditional on θ, the response random variables $X_1, X_2, \ldots, X_J$ are independently but not necessarily identically distributed. In fact, given $a_j \neq 0$, X_js are identically distributed if and only if $a_j = a_{j'}$ and $b_j = b_{j'}, \forall j \neq j'$. Therefore, the PD statistics for 2PL do not follow the exact formulation as in Cressie and Read (1984). The problem of interest here is to test a null hypothesis $H_0 : \theta = \theta_0$ against its two-tailed alternative $H_1 : \theta \neq \theta_0$. We introduce the PD family of statistics as

$$
2JI^{\lambda}(\boldsymbol{X} = \boldsymbol{x}|\theta) = \frac{2m(s - m_1)}{vs\lambda(\lambda+1)} \sum_{i=1}^{2} N_i \left(\left(\frac{N_i}{E_i} \right)^{\lambda} - 1 \right), \tag{8}
$$

where N_1 is the sum of discrimination parameters for the items that the subject answered correctly as N_2 for the incorrect items. More specifically,

$$
\begin{aligned}
N_1 &= \sum_{j=1}^{J} a_j X_j, \\
N_2 &= \sum_{j=1}^{J} a_j (1 - X_j).
\end{aligned}
$$

E_i is the expectation of N_i, e.g. $E_1 = E(N_1) = \sum_{j=1}^{J} a_j P_{\theta_0}(X_j = 1)$. Notice that $C = \frac{m_1(s-m_1)}{vs}$ is a correction factor where,

$$
m_1 = E_1/J,
$$

[1] $2nI^{\lambda}$ in its entirety is generally referred to as the PD statistic and has a Chi-squared asymptotic distribution.

$$v = var(N_1)/J = \frac{\sum_{j=1}^{J} a_j^2 P_{\theta_0}(X_j = 1)[1 - P_{\theta_0}(X_j = 1)]}{J},$$
$$s = \sum_{j=1}^{J} a_j/J.$$

From above definitions, it is clear that C does not depend on the observed response pattern $\boldsymbol{x}$. C corrects the PD statistics for independent but not necessarily identically distributed random variables so that they are asymptotically Chi-squared distributed. Similar to Cressie and Read (1984), the PD statistic is defined by continuity for $\lambda \in \{0, -1\}$, i.e.

$$2JI^0 = \lim_{\lambda \to 0} 2JI^{\lambda}(\boldsymbol{x}|\theta) = \frac{2m(s - m_1)}{vs} \sum_{i=1}^{2} N_i \log \frac{N_i}{E_i}, \tag{9}$$

for $\lambda = 0$.

The collection of statistics indexed by λ, i.e. $\{2JI^{\lambda} | \lambda \in \mathbb{R}\}$, forms the power divergence family. And we say $2JI^{\lambda}$ is the λ-power divergence statistic, or λ-PD.

2.3. *Asymptotic Equivalence of the Power Divergence Family*

In their paper, Cressie and Read (1984) proved that the PD family of statistics are asymptotically equivalent and have a Chi-squared asymptotic distribution. We derive the similar properties for the proposed PD statistics under the 2PL model, and show that the PD statistics are asymptotically equivalent and have the same limiting Chi-squared distribution with one degree of freedom.

Before proceeding to the proof, we first establish the following important lemma which will be repeatedly used later. The lemma shows that the sufficient statistic for θ under the 2PL is asymptotically normal.[2]

Lemma 1. *Let* $V_i = \frac{N_i - E_i}{E_i}$ *for* $i \in \{1, 2\}$. *Assume* a_j *is bounded and positive* $\forall j \in \mathbb{N}^+$, *and* $v \to V$ *for some* $V \in \mathbb{R}^+$ *finite, then* $\sqrt{\frac{m_1}{v}} E_1^{1/2} V_1 \xrightarrow{d} N(0, 1)$.

Proof. Let $A \in \mathbb{R}^+$ such that $a_j \leq A$, $\forall j \in \mathbb{N}^+$. By definition, $V_1 = \frac{N_1 - E_1}{E_1}$. We can rewrite $\sqrt{\frac{m_1}{v}} E_1^{1/2} V_1$ as

$$\begin{aligned}
\sqrt{\frac{m_1}{v}} E_1^{1/2} V_1 &= \sqrt{\frac{m_1}{v}} \frac{N_1 - E_1}{E_1^{1/2}}, \\
&= \sqrt{\frac{m_1}{v}} \frac{N_1 - E_1}{\sqrt{var(N_1)}} \sqrt{\frac{v}{m_1}}, \\
&= \frac{N_1 - E_1}{\sqrt{var(N_1)}}, \\
&= \frac{\sum_{j=1}^{J} \left(a_j X_j - E[a_j X_j]\right)}{\sum_{j=1}^{J} var(a_j X_j)^{1/2}}
\end{aligned} \tag{10}$$

[2] This result is perhaps well-known. But, for self-containess and rigor, a formal proof is provided here.

By Lindeberg's CLT, if we show that Lindeberg condition is met, then $\sqrt{\frac{m_1}{v}} E_1^{1/2} V_1 \xrightarrow{d} N(0, 1)$. To show Lindeberg condition, for any $\epsilon > 0$,

$$\begin{aligned}
&\lim_{J\to\infty} \frac{1}{Jv} \sum_{j=1}^{J} a_j^2 E\left((X_j - p_j(\theta_0))^2 \mathbb{1}_{|X_j - p_j(\theta_0)| > \frac{\epsilon\sqrt{Jv}}{a_j}}\right) \\
&\le \lim_{J\to\infty} \frac{1}{Jv} \sum_{j=1}^{J} a_j^2 E\left(\mathbb{1}_{|X_j - p_j(\theta_0)| > \frac{\epsilon\sqrt{Jv}}{a_j}}\right) \\
&\le \lim_{J\to\infty} \frac{1}{Jv} \sum_{j=1}^{J} a_j^2 E\left(\frac{(X_j - p_j(\theta_0))^2}{\epsilon^2 Jv/a_j^2} \mathbb{1}_{|X_j - p_j(\theta_0)| > \frac{\epsilon\sqrt{Jv}}{a_j}}\right) \\
&\le \lim_{J\to\infty} \frac{A^2}{\epsilon^2 J^2 v^2} \sum_{j=1}^{J} a_j^2 E\left((X_j - p_j(\theta_0))^2\right) \\
&\le \lim_{J\to\infty} \frac{A^2}{\epsilon^2 JV} = 0 \qquad (11)
\end{aligned}$$

□

Theorem 1. *Under correct model specification, in addition to the same conditions as in Lemma 1, assume $a_j \ge B$ and $p_j \ge \delta_p$ for some $B, \delta_p \in \mathbb{R}^+$, then $2JI^\lambda \xrightarrow{d} \chi_1^2$ as $J \to \infty$, $\forall \lambda \in \mathbb{R}$.*

Proof. By definition,

$$\begin{aligned}
2JI^1 &= \frac{m_1(s - m_1)}{vs} \sum_{i=1}^{2} E_i V_i^2, \\
&= \frac{m_1(s - m_1)}{vs} \left(E_1 V_1^2 + \frac{(N_2 - E_2)^2}{E_2}\right). \qquad (12)
\end{aligned}$$

Observing the fact that $N_1 + N_2 = \sum_{j=1}^{J} a_j = Js$, we can substitute, i.e.

$$\begin{aligned}
2JI^1 &= \frac{m_1(s - m_1)}{vs} \left(E_1 V_1^2 + \frac{(N_1 - E_1)^2}{Js - E_1}\right) \\
&= \frac{m_1(s - m_1)}{vs} \left(E_1 V_1^2 + \frac{(N_1 - E_1)^2}{E_1} \frac{E_1}{Js - E_1}\right) \\
&= \frac{m_1(s - m_1)}{vs} \left(E_1 V_1^2 + E_1 V_1^2 \frac{E_1}{Js - E_1}\right) \\
&= \frac{m_1(s - m_1)}{vs} \frac{s}{s - m_1} E_1 V_1^2 \\
&= \frac{m_1}{v} E_1 V_1^2. \qquad (13)
\end{aligned}$$

By Lemma 1 and the assumptions, $\frac{m_1}{v} E_1 V_1^2 \xrightarrow{d} Z^2$ where $Z \sim N(0, 1)$. In other words, $2JI^1$ has a asymptotic χ_1^2 distribution.

Recall $V_i = \frac{N_i - E_i}{E_i}$. Observing that $\sum_{i=1}^{2} N_i = \sum_{i=1}^{2} E_i = \sum_{j=1}^{J} a_j$, λ-PD can be expressed as

$$
\begin{aligned}
2JI^{\lambda} &= \frac{2m(s-m_1)}{vs\lambda(\lambda+1)}\sum_{i=1}^{2} N_i\left(\left(\frac{N_i}{E_i}\right)^{\lambda}-1\right)\\
&= \frac{2m_1(s-m_1)}{vs\lambda(\lambda+1)}\sum_{i=1}^{2} E_i\left((1+V_i)^{\lambda+1}-1\right)
\end{aligned}
\tag{14}
$$

We expand the right-hand side of (14) using a Taylor series with respect to V_i at the origin. It leads to

$$
\begin{aligned}
2JI^{\lambda} &= \frac{2m(s-m_1)}{vs\lambda(\lambda+1)}\sum_{i=1}^{2} E_i\left((\lambda+1)V_i+\frac{\lambda(\lambda+1)}{2}V_i^2+O(V_i^3)\right)\\
&= \frac{m_1(s-m_1)}{vs}\sum_{i=1}^{2} E_i\left(V_i^2+O(V_i^3)\right)
\end{aligned}
\tag{15}
$$

Recognizing that $2JI^1=\frac{m_1(s-m_1)}{vs}\sum_{i=1}^{2}E_iV_i^2$,

$$
2JI^{\lambda}=2JI^1+\frac{m_1(s-m_1)}{vs}\sum_{i=1}^{2}E_iO(V_i^3). \tag{16}
$$

By the strong law of large numbers for independent but not necessarily identically distributed random variables, $N_1/J-E_1/J$ converges almost surely to 0 as $J\to\infty$. Since a_j is bounded by some positive A and B, $B\delta_p\leq E_1/J\leq A$ is also bounded. It follows that $V_1=\frac{N_1/J-E_1/J}{E_1/J}\to 0$ as $J\to\infty$. Furthermore, following Lemma 1, $E_1V_1^2\xrightarrow{d}\frac{v}{m_1}Z^2$. By Slutsky's theorem $E_1V_1^3\to 0$. And, $E_2V_2^3=-(Js-E_1)\left(\frac{V_1E_1}{Js-E_1}\right)^3=-E_1V_1^3\left(\frac{E_1}{Js-E_1}\right)^2\to 0$. In addition, it can be shown that $\frac{m_1(s-m_1)}{s}$ is bounded above by $\frac{s^2}{s}=s\leq\frac{JA}{J}=A$, since $a_j\leq A$ for some A finite. It immediately follows that $2JI^{\lambda}\to 2JI^1$ as $J\to\infty$. Since we have already established that $2JI^1\xrightarrow{d}\chi_1^2$, $2JI^{\lambda}\xrightarrow{d}\chi_1^2\ \forall\lambda\in\mathbb{R}$, as $J\to\infty$. □

2.4. A Special Case

Many well-known statistics are special cases of the PD family. For example, under multinomial models, the PD statistic with $\lambda=1$ leads to the Pearson's Chi-squared statistic (Cressie & Read, 1984). For these models, Pearson's Chi-squared statistic coincides with the test statistic used in Rao's score test (Agresti, 2011). In our case, the PD statistic with $\lambda=1$ is equivalent to the score test statistic of the person parameter under the 2PL.

For testing the simple null hypothesis $H_0:\theta=\theta_0$, the score test statistic is given by

$$
S(\theta_0)=\frac{U(\theta_0)^2}{I(\theta_0)}, \tag{17}
$$

where $U(\theta_0)=\frac{\partial\log L(\theta|\boldsymbol{x})}{\partial\theta}|_{\theta=\theta_0}$ is the score, and $I(\theta_0)=-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{x})|_{\theta=\theta_0}$ is the Fisher information at θ_0. Using the likelihood specified in Eq. 2, the score test statistic for θ under 2PL is

$$S(\theta_0) = \frac{\left(\sum_{j=1}^{J} a_j x_j - \sum_{j=1}^{J} a_j p_j(\theta_0)\right)^2}{\sum_{j=1}^{J} a_j^2 p_j(\theta_0)[1 - p_j(\theta_0)]}. \tag{18}$$

When $\lambda = 1$, by Eq. 13, the PD statistic can be written as

$$\begin{aligned} 2JI^1 &= \frac{m_1}{v} E_1 V_1^2 = \frac{E_1}{var(N_1)} \frac{E_1(N_1 - E_1)^2}{E_1^2} \\ &= \frac{(N_1 - E_1)^2}{var(N_1)} = \frac{\left(\sum_{j=1}^{J} a_j x_j - \sum_{j=1}^{J} a_j p_j(\theta_0)\right)^2}{\sum_{j=1}^{J} a_j^2 p_j(\theta_0)[1 - p_j(\theta_0)]}, \end{aligned} \tag{19}$$

which coincides with the score test statistic as in Eq. 18.

2.5. *A Moment Matching Method for Choosing λ*

The proposed PD family of statistics are asymptotically equivalent and have the same limiting χ^2 distribution with one degree of freedom. However, under a given finite sample size, the error of approximating the exact probability distribution of the PD statistic using the asymptotic χ_1^2 distribution might be different for different λs. Similar to Cressie and Read (1984), we approximate the exact first moment of the PD family $2JI^\lambda$ by a second-order expansion. Then we choose the optimal λs based on matching the approximated exact first moment to that of the χ_1^2 distribution.

We first approximate the exact first moment of the PD family of statistics. Let $W_i = (N_i - E_i)/\sqrt{J}$, and remember that $V_i = (N_i - E_i)/E_i$ and $C = [m_1(s - m_1)]/(vs)$. $2JI^\lambda$ can be expanded as

$$\begin{aligned} 2JI^\lambda &= \frac{2C}{\lambda(\lambda+1)} \sum_{i=1}^{2} E_i\{(\lambda+1)V_i + \frac{\lambda(\lambda+1)}{2} V_i^2 + \frac{\lambda(\lambda+1)(\lambda-1)}{6} V_i^3 \\ &\quad + \frac{\lambda(\lambda+1)(\lambda-1)(\lambda-2)}{24} V_i^4 + O(V_i^5)\} \\ &= C\left(\sum_{i=1}^{2} E_i V_i^2 + \frac{\lambda-1}{3} \sum_{i=1}^{2} E_i V_i^3 + \frac{(\lambda-1)(\lambda-2)}{12} \sum_{i=1}^{2} E_i V_i^4 + \sum_{i=1}^{2} O(E_i V_i^5)\right). \end{aligned} \tag{20}$$

Notice that $\sum_{i=1}^{2} O(E_i V_i^5) = O(V_1^3) = O(J^{-3/2})$. Therefore,

$$2JI^\lambda = C\left(\sum_{i=1}^{2} \frac{W_i^2}{m_i} + \frac{\lambda-1}{3\sqrt{J}} \sum_{i=1}^{2} \frac{W_i^3}{m_i^2} + \frac{(\lambda-1)(\lambda-2)}{12J} \sum_{i=1}^{2} \frac{W_i^4}{m_i^3} + O(J^{-3/2})\right). \tag{21}$$

In order to get an approximation of the first moment of the exact distribution of the PD family of statistics, we need to take expectation of the approximated $2JI^\lambda$ as in Eq. 21. We treat each sums separately. The first sum can be simplified as

$$C \sum_{i=1}^{2} \frac{W_i^2}{m_i} = C\left(\frac{W_1^2}{m_1} + \frac{W_1^2}{s - m_1}\right) = \frac{m_1(s - m_1)}{vs} \frac{s}{m_1(s - m_1)} W_1^2 = \frac{W_1^2}{v}. \tag{22}$$

So its expectation is

$$E\left(C\sum_{i=1}^{2}\frac{W_i^2}{m_i}\right)=E\left(\frac{W_1^2}{v}\right)=\frac{1}{Jv}E\left((N_1-E_1)^2\right)=1. \tag{23}$$

The second sum is

$$\begin{aligned}C\sum_{i=1}^{2}\frac{W_i^3}{m_i^2}&=C\left(\frac{W_1^3}{m_1^2}-\frac{W_1^3}{(s-m_1)^2}\right)=C\frac{s(s-2m)}{m_1^2(s-m_1)^2}W_1^3\\&=C\frac{s(s-2m)}{J^{3/2}m_1^2(s-m_1)^2}(N_1-E_1)^3.\end{aligned} \tag{24}$$

And the third sum is

$$\begin{aligned}C\sum_{i=1}^{2}\frac{W_i^4}{m_i^3}&=C\left(\frac{1}{m_1^3}+\frac{1}{(s-m_1)^3}\right)W_1^4\\&=C\frac{s(s^2-3ms+3m^2)}{J^2m^3(s-m_1)^3}(N_1-E_1)^4.\end{aligned} \tag{25}$$

Then, the expectation of $2JI^{\lambda}$ becomes

$$\begin{aligned}E(2JI^{\lambda})&=E\left(C\sum_{i=1}^{2}\frac{W_i^2}{m_i}\right)+E\left(C\frac{\lambda-1}{3\sqrt{J}}\sum_{i=1}^{2}\frac{W_i^3}{m_i^2}+C\frac{(\lambda-1)(\lambda-2)}{12J}\sum_{i=1}^{2}\frac{W_i^4}{m_i^3}\right)+O(J^{-3/2})\\&=1+\frac{C(\lambda-1)s}{3Jm^2(s-m_1)^2}\{\frac{(s-2m)E[(N_1-E_1)^3]}{J}\\&\quad+\frac{(\lambda-2)(s^2-3ms+3m^2)}{4J^2m(s-m_1)}E[(N_1-E_1)^4]\}+O(J^{-3/2}).\end{aligned} \tag{26}$$

Comparing the approximated first moment of the exact distribution of $2JI^{\lambda}$ in Eq. 26 to that of the χ_1^2 distribution, we find that $E\left(C\sum_{i=1}^{2}\frac{W_i^2}{m_i}\right)=1$ matches exactly. In other words, the size of $E(C\sum_{i=1}^{2}\frac{W_i^3}{m_i^2}+C\sum_{i=1}^{2}\frac{W_i^4}{m_i^3})$ gives some information about the error of using the limiting χ_1^2 distribution to approximate the exact distribution of $2JI^{\lambda}$. Naturally we want to choose λ such that it minimizes the correction term

$$\begin{aligned}&E\left(C\frac{\lambda-1}{3\sqrt{J}}\sum_{i=1}^{2}\frac{W_i^3}{m_i^2}+C\frac{(\lambda-1)(\lambda-2)}{12J}\sum_{i=1}^{2}\frac{W_i^4}{m_i^3}\right)\\&=\frac{C(\lambda-1)s}{3Jm^2(s-m_1)^2}\{\frac{(s-2m)E[(N_1-E_1)^3]}{J}\\&\quad+\frac{(\lambda-2)(s^2-3ms+3m^2)}{4J^2m(s-m_1)}E[(N_1-E_1)^4]\}\\&=0\end{aligned} \tag{27}$$

Solve Eq. 27 for λ. Clearly, $\lambda_1 = 1$ is an obvious solution and gives the null values of all the expanded terms beyond the first one. The other solution is

$$\lambda_2 = 2 - \frac{4(E[(N_1 - E_1)^3]/J)m_1(s - m_1)(s - 2m)}{(E[(N_1 - E_1)^4]/J^2)(s^2 - 3ms + 3m^2)}, \tag{28}$$

where

$$E[(N_1 - E_1)^3] = \sum_{j=1}^{J} a_j^3 p_j(1 - p_j)(1 - 2p_j), \tag{29}$$

and

$$E[(N_1 - E_1)^4] = 3\sum_{i \neq j} a_i^2 a_j^2 p_i(1 - p_i)p_j(1 - p_j) + \sum_{j=1}^{J} a_j^4 p_j(1 - p_j)(1 - 3p_j + 3p_j^2). \tag{30}$$

Note that p_j is the probability of a correct response to jth item under the null hypothesis $H_0 : \theta = \theta_0$, i.e. $p_j = P_{\theta_0}(X_j = 1)$. Therefore, λ_2 is a function of item parameters and θ_0. We simulated these parameters from their usual distributions, i.e. $a_j \sim \text{unif}(0.5, 2.0)$, $b_j \sim N(0, 1)$, and $\theta \sim N(0, 1)$, and computed the values of λ_2. Figure 1 shows the distributions of λ_2 based on a simulation of 10,000 times for 1PL and 2PL under different numbers of items. λ_2 is usually in the range between 1 and 2. With a smaller number of items, λ_2 is more likely around 2 but still has considerable frequencies across the entire range. However, as the number of items increases, the distribution of λ_2 tends to concentrate more and more its density around 2. This observation is interesting as Cressie and Read (1984) shows that both $\lambda = 1$ and $\lambda = 2$ minimizes the correction term in their expansion of the first moment of the PD statistic for binomial models when $\sum_{i=1}^{2}(1/\pi_{0i}) = 4$ or $\pi_{01} = \pi_{02} = 0.5$.

To examine how its distribution relates to θ, we plot the distribution against different θs under usual item parameter distributions in Fig. 2.

At $\theta = 0$, the center of the distribution of λ_2 is very close to 2. This observation agrees with Cressie and Read (1984). At $\theta = 0$, the probability of a correct response is most likely to be 0.5 as the item difficulty parameter is sampled from a standard normal distribution. The center moves toward smaller values as θ moves away from 0 in either direction. This trend becomes more obvious with more items.

2.6. Hypothesis Testing and Confidence Sets

Using its limiting distribution as an approximation, under the null hypothesis $H_0 : \theta = \theta_0$, the PD statistic $2JI^{\lambda}$ is distributed according to a Chi-squared distribution with one degree of freedom. Utilizing this result, we formalize the hypothesis testing procedure using PD statistics and the associated confidence sets.

Consider testing the simple null hypothesis $H_0 : \theta = \theta_0$ versus the two-tailed alternative $H_1 : \theta \neq \theta_0$ with a test

$$\phi = \begin{cases} 1 & \text{if } 2JI^{\lambda}(\boldsymbol{x}) \geq \kappa_{\alpha/2}^2 \\ 0 & \text{if } 2JI^{\lambda}(\boldsymbol{x}) < \kappa_{\alpha/2}^2, \end{cases}$$

where $\kappa_{\alpha/2}^2$ is the $(1-\alpha)$th percentile of the χ_1^2 distribution. If $\phi = 1$, the test rejects H_0. Otherwise, the test fails to reject H_0. Assuming that J is sufficiently large, $P_{\theta_0}(2JI^{\lambda}(\boldsymbol{X}) \geq \kappa_{\alpha/2}^2) \approx \alpha$, for

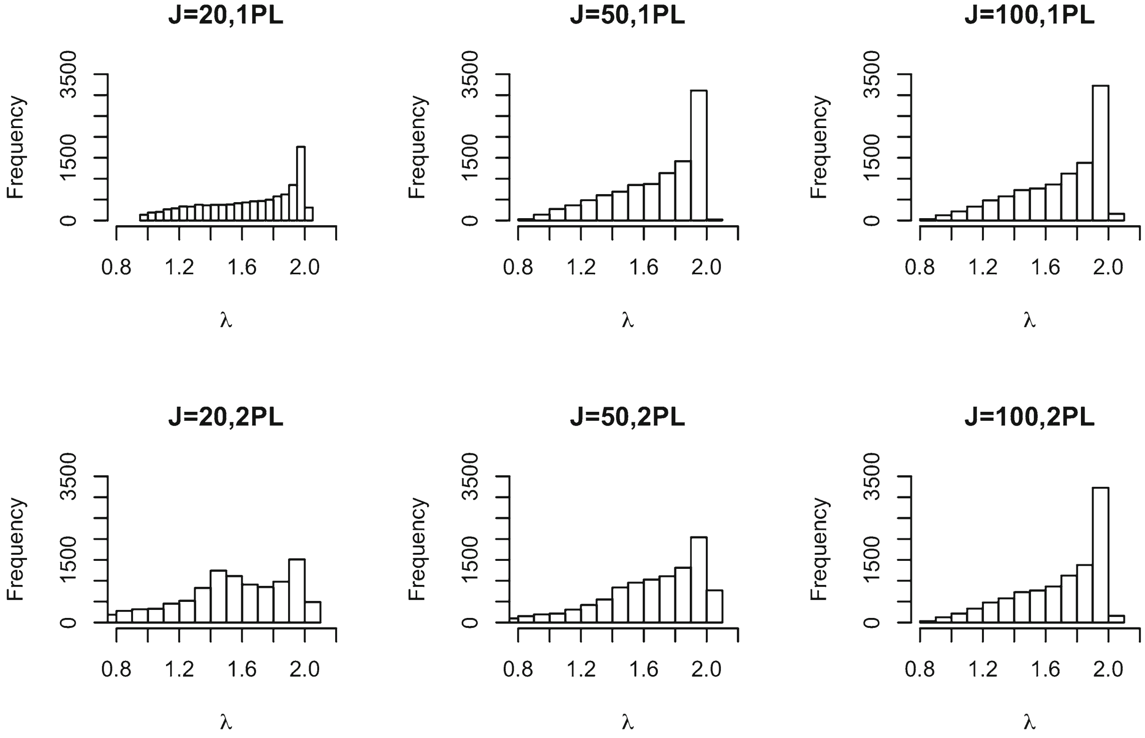


FIGURE 1.
Distributions of λ_2 under typical model parameters

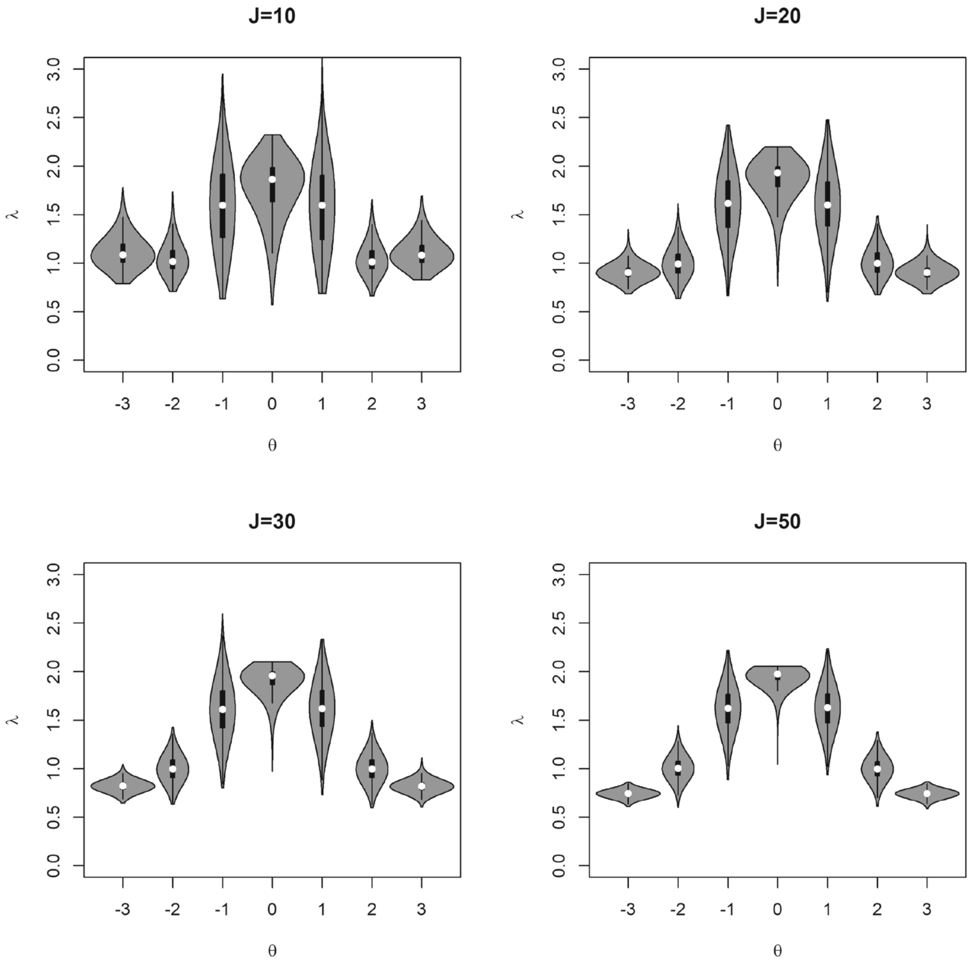


FIGURE 2.
Distributions of λ_2 against θ

some nominal level α. In general, due to the approximation error and the discreteness of the PD statistic under IRT models, the *size* of the test ϕ, $\alpha^\star = P_{\theta_0}(2JI^\lambda(\boldsymbol{X}) \geq \kappa^2_{\alpha/2})$ does not equal to the nominal level α exactly.

The confidence set of θ can be obtained by inverting the two-tailed test ϕ. Given a response pattern $\boldsymbol{X} = \boldsymbol{x}$, we find all θ under which the null hypothesis will be not be rejected, i.e.

$$A_\theta(\boldsymbol{x}) = \{\theta : \phi_\theta(\boldsymbol{x}) = 0\}. \tag{31}$$

The confidence set obtained by inverting a hypothesis test generally is not necessarily an interval. Under some circumstances, the confidence set is guaranteed to be an interval provided the end points of the acceptance region is monotone in θ_0 (e.g. Liu et al., 2018). In the context of the binomial models, Jin, Thulin, and Larsson (2017) proved that inverting the test using Cressie and Read's PD statistics leads to the confidence interval if $\lambda \in [0, 1]$. However, their results cannot be easily extended to the current problem. In fact, we do observe confidence sets consisting of two disconnected intervals in rare occasions. That being said, the confidence set $A(\boldsymbol{x})$ is an interval in most cases based on our simulations. In case of the existence of the disconnected intervals, one viable approach is to find the minimal interval that contains the disconnected intervals (Doebler et al., 2012), i.e.

$$I_\theta(\boldsymbol{x}) = \{\theta : l_\theta(\boldsymbol{x}) \leq \theta \leq u_\theta(\boldsymbol{x})\}, \tag{32}$$

where $l_\theta(\boldsymbol{x}) = \inf A_\theta(\boldsymbol{x})$, and $u_\theta(\boldsymbol{x}) = \sup A_\theta(\boldsymbol{x})$. Computationally, we can find the set $A(\boldsymbol{x})$ by brute forcing over a finite grid of θ. This is feasible as the PD statistic has a closed form and time required for each evaluation is minimal. In this paper, we chose the grid to be from -6.0 to 6.0 by increment of 0.01. This aligns with the common practice that the estimates of θ are reported with a precision of two decimal places (Doebler et al., 2012).

It is well-understood that the MLE of the person parameter is unbounded when a perfect response pattern is observed (i.e. all zeros or all ones). Interval estimation methods that depend on the existence of a finite estimate of the person parameter, such as the saddlepoint approximation, generally cannot apply to perfect response patterns. The current approach, on the other hand, obtains the confidence interval of the person parameter by inverting the λ-PD test. Under perfect response patterns, the λ-PD, $2JI^\lambda(\theta)$, is observed to be monotone in θ.[3] Consequently, the resulting confidence interval is unbounded on one side in these cases. If a bounded interval is desired for any reason, we recommend truncating the interval at some large boundary, for example, -6.0 and 6.0. As ranges beyond these limits are rarely of practical interest.

3. Simulations

3.1. Comparisons of the Distribution of PD Statistics to χ^2_1 Distribution

A small simulation was performed to investigate the level of discrepancies between the exact distribution of the PD statistics and the theoretical asymptotic χ^2_1 distribution. We considered two sample sizes: $J = 10$ items and $J = 50$ items. Additionally, for each sample size, we also compared $\lambda = 1, \lambda = \lambda_2$, and $\lambda = 4$. Item parameters and level of latent trait are treated as random effects here. Thus, for each replication, item parameters and the latent variable are generated from their usual distributions, i.e. $a_j \sim unif(0.5, 2.0)$, $b_j \sim N(0, 1)$ for $j = 1, 2, \ldots, J$, and $\theta \sim N(0, 1)$. Then a vector of J binary responses is generated according to the 2PL. Based on the response pattern and item parameters, 1-PD, λ_2-PD, and 4-PD are computed. For each J, the process is replicated 50,000 times.

Figure 3 shows the resulting quantiles of the PD statistics against the quantiles of the χ^2_1 distribution under different sample sizes. When $J = 10$, the distribution of the PD statistics is reasonably close to the χ^2_1 distribution for $\lambda = 1$ and $\lambda = \lambda_2$, and $\lambda = 1$ seems to have a slightly better approximation. But the distributions do deviate noticeably from the χ^2_1 distribution in the tail. Meanwhile, for $\lambda = 4$, the difference is significant almost across the entire support. With an increased number of items, $J = 50$, the distributions of the PD statistics become closer to the asymptotic distribution. In fact, when $\lambda = 1$ or $\lambda = \lambda_2$, the distribution of the statistics shows very little deviance to the χ^2_1 distribution even for the tail region. Though showing an improvement over the smaller sample size, the distribution of 4-PD still differs significantly from the χ^2_1 distribution. This result is consistent with the findings from the moment matching method introduced earlier where we have shown that both $\lambda = 1$ and $\lambda = \lambda_2$ minimize the correction term.

3.2. Confidence Interval Coverage Rate and Length

Based on the theoretical analysis as well as the numerical demonstration, 1-PD and λ_2-PD seem to converge faster to their limiting χ^2_1 distribution compared to other members of the PD

[3] One reviewer remarked correctly that it is straightforward to mathematically show the monotonicity assuming $a_j = 1$ and $b_j = b\,\forall j$; However, it is non trivial to prove under the general case.

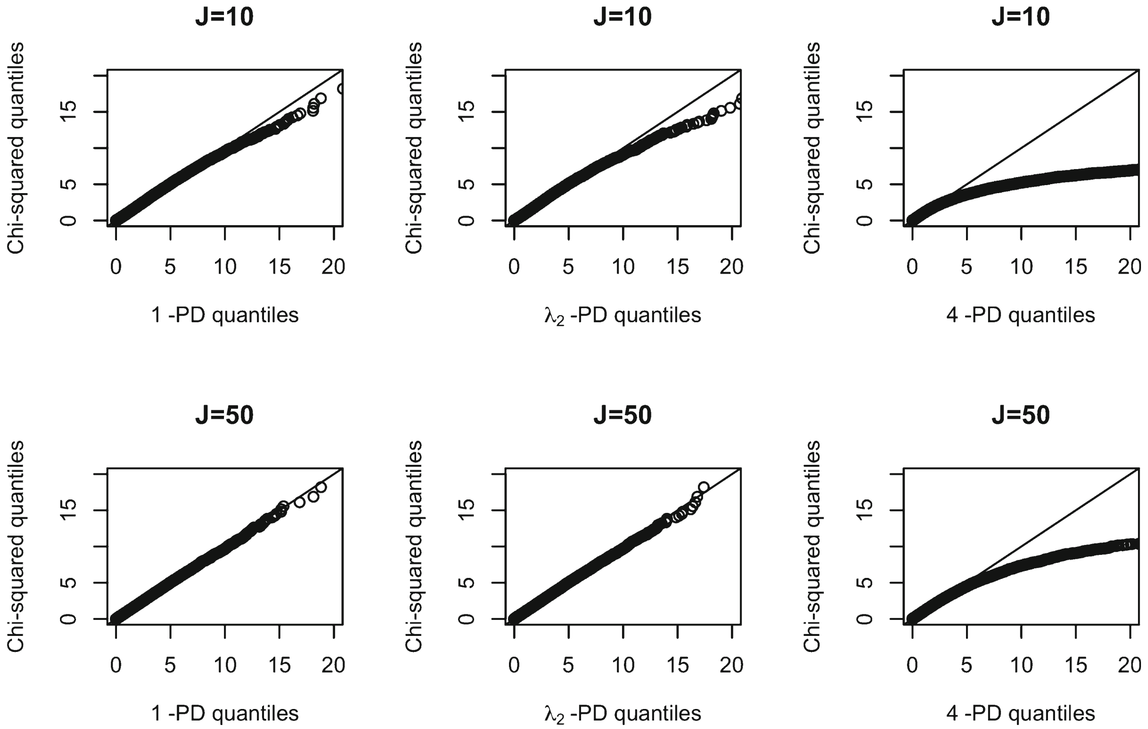


FIGURE 3.
χ_1^2 quantile plots for the PD statistics

family. Naturally, it is reasonable to expect that the confidence intervals by inverting the hypothesis tests with these choices of λ may have good coverage rates that are closer to the nominal level. In this simulation, we examine the performance of the proposed PD confidence intervals under small to moderate sample sizes. In addition, the results are benchmarked against the saddlepoint approximation method (Biehler et al., 2014) which has been shown to result in some highly accurate confidence intervals under small sample sizes.

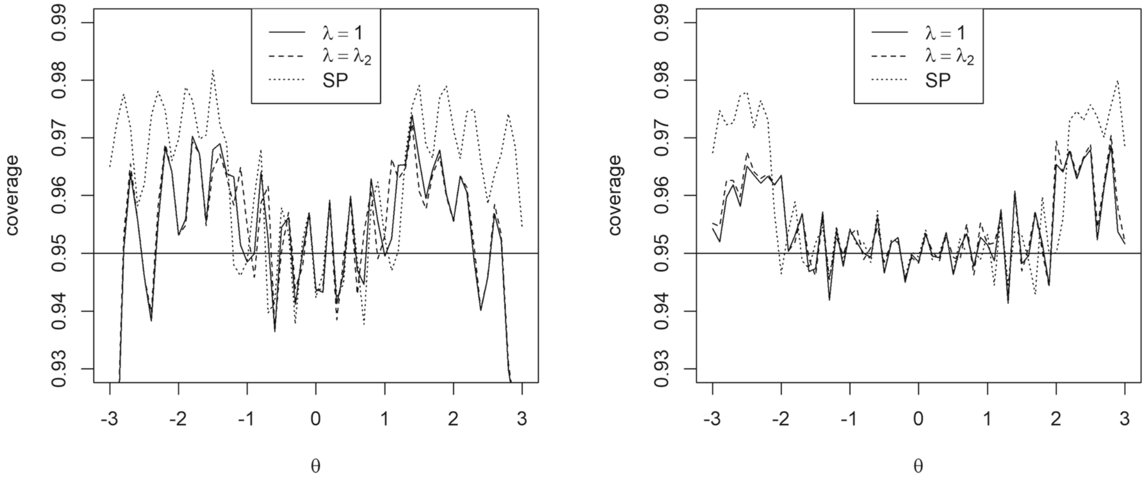
The simulation conditions considered here are similar to those of Biehler et al., (2014). Under the 2PL model, the item difficulty parameters are given by $b_j = \Phi^{-1}((j-0.5)/J)$ for $j = 1, 2, \ldots, J$ where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard Gaussian distribution. Item discrimination parameters a_js are randomly generated from Lognormal$(0, 0.15^2)$. We considered a small sample size $J = 15$ and a medium sample size $J = 30$. For each sample size, the true latent trait θ is taken over the one-dimensional grid from -3.0 to 3.0 by step of 0.1. At each latent trait level, we generated 10,000 response patterns, and perfect patterns having all 1s or all 0s were excluded. For each response pattern, we compute 1-PD confidence interval, λ_2-PD confidence interval, and the confidence interval based on the saddlepoint approximation using the Lugannai–Rice formula (Lugannani & Rice, 1980) all at the 0.95 nominal coverage probability.

The expected coverage rates across the latent trait levels are illustrated in Fig. 4.

Regardless the choice of λ, the two PD confidence intervals perform almost identically across different latent trait levels and provide coverage close to the nominal level in the medium range of the latent trait where the items have adequate information about θ. As θ moves toward either extreme, the coverage rate tends to inflate. The saddlepoint approximation-based confidence interval (SP confidence interval) provides very comparable coverage in the middle, but tends to inflate more in the extremes. We note that the oscillation in the expected coverage rates is due to the discreteness of the item response sample space which has also been remarked in previous research (e.g. Biehler et al., 2014). The amplitude of the oscillation decreases with a larger number of items as the number of possible observed response patterns increases.

Compared to other approaches, such as the Wald confidence intervals and the asymptotic cumulants of the studentized person parameter (Ogasawara, 2012), Biehler et al. (2014) show that the SP confidence interval has favorable performance under the same simulation conditions considered in the current simulation. The current results suggest the 1-PD confidence interval and the λ_2-PD confidence interval perform at least as well as the SP confidence interval in the medium latent trait range while possibly provide coverage rates closer to the nominal level in the extremes. For a $1-\alpha$ nominal coverage probability, the two-sided SP confidence interval is obtained by using two one-sided SP intervals, each providing $1-\alpha/2$ coverage in one direction. Under a finite sample size, for sufficiently small and sufficiently large θ, the coverage probability is actually around $1-\alpha/2$ rather than the nominal $1-\alpha$ using this approach (see Agresti, 2003 for a similar discussion). For example, under the 2PL, there exists some θ_0 sufficiently small so that $P_{\theta_0}(\boldsymbol{X} = \boldsymbol{x}_0) \geq \alpha/2$ even for $\boldsymbol{x}_0 = (0, 0, \ldots, 0)$. Then θ_0 will be covered by the one-sided confidence interval in the direction $\theta \leq \theta_0$ with a probability of 1 regardless of the observed pattern. Consequently, θ_0 can only be possibly be excluded in the other direction with a probability of $\alpha/2$. Thus, it leads to a two-sided confidence interval with a coverage probability of $1-\alpha/2$ instead of the intended $1-\alpha$. This limitation of using two one-sided intervals is reflected in Fig. 4. The PD confidence interval, on the other hand, is obtained by inverting a two-sided test. As a result, it is generally less prone to the inflation of the coverage probability in the extremes.

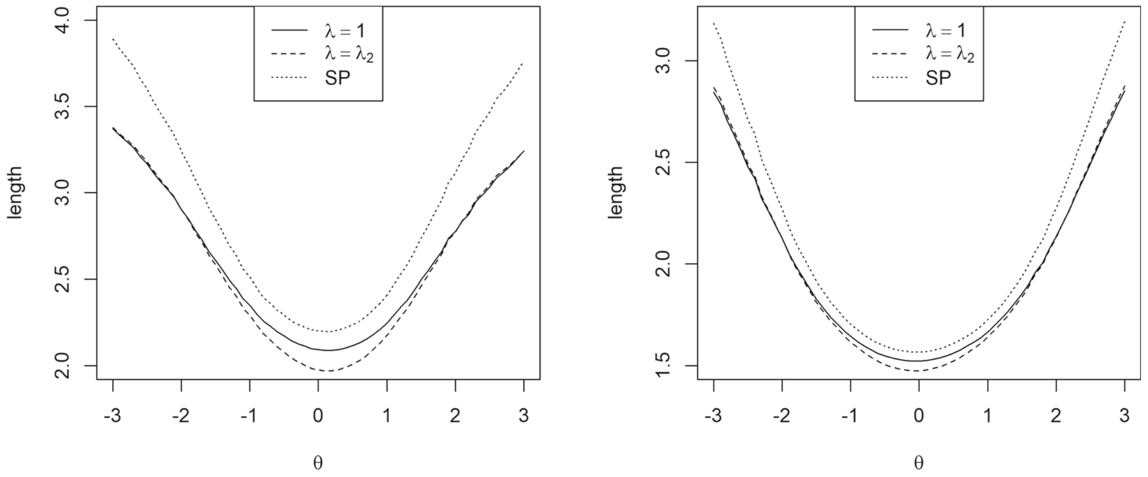
Figure 5 shows the expected confidence interval length over the 10,000 replications. On average, the confidence intervals are shorter in the middle and much longer in the extremes since the item parameters are generated with more information about θ in the middle but little information about extreme θ. As the number of items increases, the average lengths decrease. While all three confidence intervals provide coverage close to the nominal level in the middle, the two PD confidence intervals seem to be uniformly shorter than the SP confidence interval. The



(a) 15 items

(b) 30 items

FIGURE 4.
Expected coverage rates



(a) 15 items

(b) 30 items

FIGURE 5.
Expected confidence interval length

difference is more obvious in the extreme θ where the coverage probability of the SP confidence interval is more inflated. Moreover, choosing $\lambda = \lambda_2$ leads to typically shorter confidence intervals than those from $\lambda = 1$ especially for medium θ despite the identical coverage probabilities.

3.3. Coverage Under Model Misspecification

The methods developed so far are based on the assumption that the IRT model is correct. However, in real world, the model could be misspecified for many reasons, such as violations of uni-dimensionality or local independence, treating estimated item parameters as true population values, etc. While the asymptotic theory based on the correct model specification serves an useful approximation of the reality, it is important to understand its limitation. In this section, we examine the coverage rate of the PD confidence intervals under model misspecification.

We follow Ogasawara (2012, 2013) for simulating model misspecification cases. Let P_{Tj} denote the true item response probability to the jth item for $j = 1, 2, \dots, J$. The IRT model is misspecified if at least one item response probability under the 2PL is different from the true item response probability, i.e. $P_j \neq P_{Tj}$ for some j. For the true item response probabilities, we assume the local independence assumption still holds. Thus, the joint probability of observing a response vector can be factorized, $P_T(\boldsymbol{X} = \boldsymbol{x}) = \prod_j P_{Tj}(x_j)$. Under the misspecification, the population θ under the 2PL is defined as θ that satisfies the first-order condition of the maximum likelihood estimation in the population, i.e.

$$J E_T(\frac{\partial \bar{l}}{\partial \theta}) = \sum_{j=1}^{J} \left(\frac{P_{Tj}}{P_j} \frac{\partial P_j}{\partial \theta} + \frac{1 - P_{Tj}}{1 - P_j} \frac{\partial (1 - P_j)}{\partial \theta} \right) = 0. \tag{33}$$

Intuitively, the population θ minimizes the Kullback–Leibler (KL) divergence between the true model and the 2PL, $\theta := \text{argmin}_\theta D_{KL}(P_T | P)$. The true item response probability to the jth item is given by

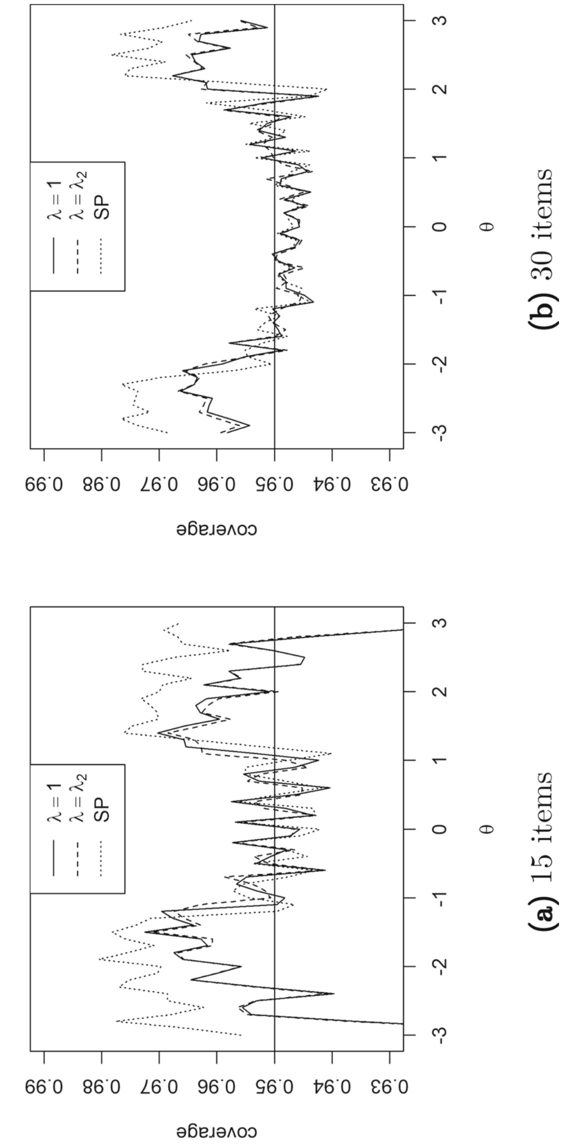
$$P_{Tj} = \frac{1}{1 + \exp\left(-a_j(\theta - b_j) + e_j + e_\theta\right)}, \tag{34}$$

where $e_j \sim N(0, \sigma_{e_j}^2)$ with $\sigma_{e_j}^2 = 0.1$ and 1.0 corresponding to slight and gross model misspecification. Parameters a_j and b_j are generated in the same way as before, $a_j \sim \text{lognormal}(0, 0.15^2)$ and $b_j = \Phi^{-1}((j - 0.5)/J)$ for $j = 1, 2, \dots, J$. e_θ is then numerically solved for a grid $\theta = -3.0, -2.9, \dots, 3.0$ so that Eq. 33 holds. Item responses are then generated using the true model. Perfect response patterns are discarded. Coverage rates of the PD confidence intervals as well SP confidence intervals are examined. Results based on 10,000 repetitions are in Figs. 6 and 7.

While the coverage rates are not too far away from the nominal level under slight model misspecification for medium level of θ, gross model misspecification results in inflated coverage especially for smaller number of items.

4. A Real Data Example

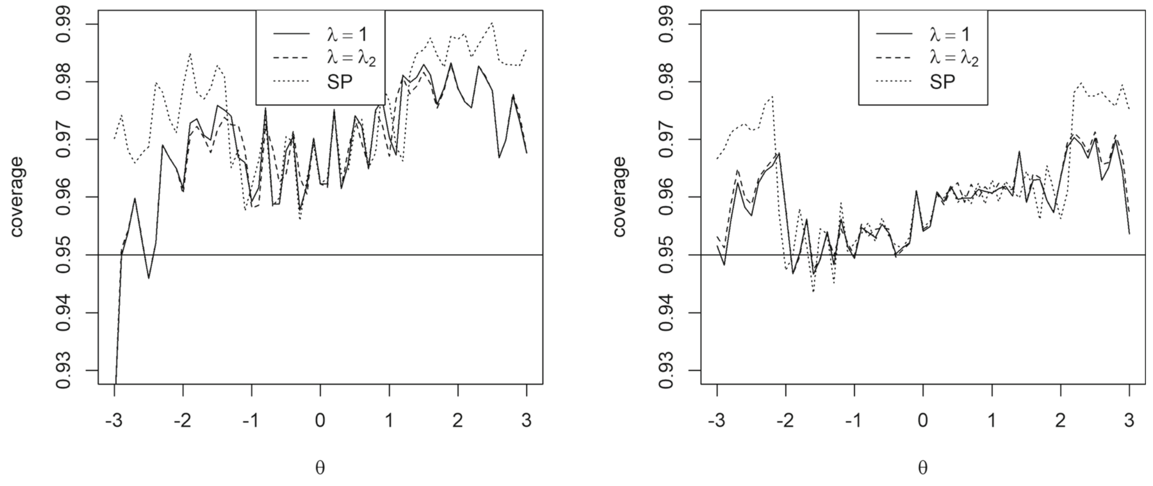
To demonstrate the utility of the proposed PD statistics for purpose of the interval estimation of the person parameter, we analyze the LSAT dataset. The LSAT dataset consists of dichotomous responses from 1000 individuals to 5 questions measuring a single latent ability. The dataset can be accessed through the R package ltm (Rizopoulos, 2015). We fitted the 1PL model with item discrimination parameters constrained to be the same across all items, i.e. $a_j = a, \ \forall j$. The item parameter estimates and the associated standard errors are in Table 1.



(a) 15 items

(b) 30 items

FIGURE 6.
Coverage rates under slight model misspecification



(a) 15 items

(b) 30 items

FIGURE 7.
Coverage rates under gross model misspecification

Table 1.
Item parameter estimates and fit statistics

	Difficulty	SE	χ^2	*P* value
Item 1	− 3.615	0.327	61.929	0.515
Item 2	− 1.322	0.142	159.010	0.584
Item 3	− 0.318	0.098	233.786	0.327
Item 4	− 1.730	0.169	132.473	0.663
Item 5	− 2.780	0.251	84.760	0.683
Dscrmn	0.755	0.069		

P values are bootstrapped rather than from the asymptotic distribution.

The standard errors for the parameter estimates are smaller for the medium level difficulty items and larger for the more extreme items. For the purpose of the demonstration of the proposed method, the estimated item parameters are treated as fixed in this example.

To evaluate the overall goodness of fit, we compute a Pearson's Chi-square statistic,

$$T = \sum_{\boldsymbol{x} \in S} \frac{(O(\boldsymbol{x}) - E(\boldsymbol{x}))^2}{E(\boldsymbol{x})}, \tag{35}$$

where S denotes the set of all 2^J possible response patterns, $O(\boldsymbol{x})$ is the observed frequency of response pattern $\boldsymbol{x}$, and $E(\boldsymbol{x})$ is the expected frequency. We bootstrapped the distribution of the statistic T using the estimated standard errors of the item parameter estimates. This approach has been implemented in the widely used ltm R package. Based on 100 bootstrap samples, the *P* value is 0.812. The overall fit of the Rasch model is good. The item level fit of the five items is also adequate (see Table 1).

Since the raw sum score is sufficient for estimating θ, we compute the 1-PD confidence interval for each of the 6 raw sum scores. Due to the small number of items, computing confidence intervals based on the exact distribution of the sufficient statistics (exact CIs) is feasible (see Klauer, 1991; Liu et al., 2018 for details). Exact CIs are included for comparison. The lower and upper limits of the intervals are plotted in Fig. 8.

The PD confidence intervals are uniformly shorter than the exact confidence intervals. However, their potential coverage could be very different under such a small number of items. To examine the potential coverage probabilities of the confidence intervals, we use a similar approach used by Liu, Hannig, and Pal Majumder (2019). For a grid of θ from -3.0 to 3.0 by 0.1, the expected coverage probabilities are computed by

$$P(\theta \in I_\theta|\theta) = \sum_{\boldsymbol{x}} \mathbb{1}(\theta \in I_\theta(\boldsymbol{x}))P(\boldsymbol{x}|\theta). \tag{36}$$

The results are in Fig. 9.

The coverage probabilities of the PD confidence interval oscillate around the nominal level; while the exact confidence interval shows inflated coverage.

5. Discussion

In this paper, we proposed a PD family of statistics for testing the person parameter in IRT models which includes the score test as a special case with $\lambda = 1$. Although the PD family of

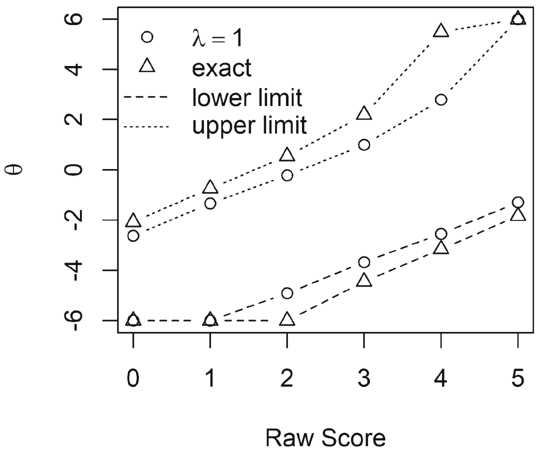


FIGURE 8.
95% CI limits for the LSAT example

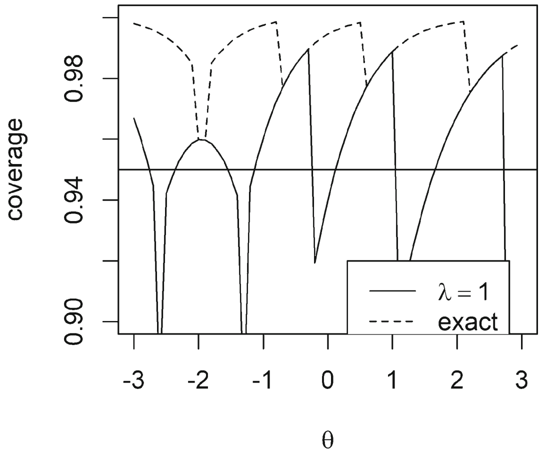


FIGURE 9.
Expected coverage probabilities for the LSAT example

statistics are asymptotically equivalent and have the same limiting distribution, for short to medium test lengths, 1-PD and λ$_2$-PD seem to be the optimal choices. It is theoretically justified that they minimizes the correction term in matching the first moment of the distribution of the statistics to that of the reference χ_1^2 distribution. In addition, the optimality is demonstrated empirically through simulations.

In their original paper, Cressie and Read (1984) used the term "goodness-of-fit" tests when introducing the PD family of statistics for the multinomial models. Readers should not confuse the current PD statistics for the person parameters with the person fit statistics. Conditional on the estimated latent trait, a person fit statistic measures the discrepancies between the observed response pattern and what the model implies. The most popular person fit statistic is perhaps the likelihood-based l_z statistic and its corrected version l_z^* (e.g. Drasgow, Levine, & Williams, 1985; Meijer & Sijtsma, 2001; Snijders, 2001; Sinharay, 2016). The proposed PD family of statistics, on the other hand, tests for the value of θ given the observed response pattern and the model.

Compared to the exact distribution approach proposed for testing θ, the current approach is much less computationally intensive. But at the same time, it also shows good small sample properties which makes it a viable and attractive option for the problem of interval estimation of the latent trait.

The current paper introduces a method for generalizing the results of Cressie and Read (1984) to cases where multinomial random variables are independently but not necessarily identically distributed. Although the PD family of statistics in this paper are developed under the 2PL model, the method can potentially be extended to a wide class of IRT models with additional work. For example, the multidimensional 2PL model, where the sufficient statistic for the multivariate person parameter is a natural multidimensional extension of the 2PL results, may have a multivariate PD family of statistics under similar regularity conditions. Potential extensions may also include certain types of polytomous IRT models. However, some other IRT models, such as the 3PL, do not have a simple one-dimensional sufficient statistic for the person parameter. Extending current method of developing PD statistics for these models may not be straightforward.

Developing procedures for interval estimation of the ability parameter that are robust against model misspecification should also be considered in the future. Ogasawara (2019) developed robust asymptotic standard errors (ASEs) of the minimum ϕ-divergence estimators for the multinomial parameters. However, it cannot be directly used in the context of interval estimation of the person parameter. It would be useful to develop a sandwich type of ASE (e.g. Ogasawara, 2013) of distributions of power divergence statistics.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Agresti, A. (2003). Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research*, *12*(1), 3–21.

Agresti, A. (2011). *Score and pseudo-score confidence intervals for categorical data analysis* (Vol. 3(2)). London: Taylor & Francis. https://doi.org/10.1198/sbr.2010.09053.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press.

Basu, A., Shioya, H., & Park, C. (2011). *Statistical inference: The minimum distance approach*. London: Taylor & Francis. https://doi.org/10.1080/02664763.2012.681565.

Biehler, M., Holling, H., & Doebler, P. (2014). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika*, *80*(3), 665–688. https://doi.org/10.1007/s11336-014-9405-1.

Casella, G., & Berger, R. (2001). *Statistical inference*. New York: Duxbury Resource Center. Textbook Binding.

Cressie, N. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(3), 440–464. https://doi.org/10.2307/2345686.

Doebler, A., Doebler, P., & Holling, H. (2012). Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika*, *78*(1), 98–115. https://doi.org/10.1007/s11336-012-9290-4.

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*,. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x.

Felipe, A., Miranda, P., & Pardo, L. (2015). Minimum Φ-divergence estimation in constrained latent class models for binary data. *Psychometrika*,. https://doi.org/10.1007/s11336-015-9450-4.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing. https://doi.org/10.1007/978-94-017-1988-9.

Harel, D. (2014). *The effect of model misspecification for polytomous logistic adjacent category item response theory models*. Unpublished doctoral dissertation. Montreal, QC: McGill University.

Jin, S., Thulin, M., & Larsson, R. (2017). Approximate bayesianity of frequentist confidence intervals for a binomial proportion. *American Statistician*, *71*(2), 106–111. https://doi.org/10.1080/00031305.2016.1208630.

Klauer, K. C. (1991). Exact and best confidence intervals for the ability parameter of the Rasch model. *Psychometrika*, *56*(3), 535–547. https://doi.org/10.1007/BF02294489.

Liu, X., Han, Z., & Johnson, M. S. (2018). The UMP exact test and the confidence interval for person parameters in IRT models. *Psychometrika*, *83*(1), 182–202. https://doi.org/10.1007/s11336-017-9580-y.

Liu, Y., Hannig, J., & Pal Majumder, A. (2019). Second-order probability matching priors for the person parameter in unidimensional IRT models. *Psychometrika*, *84*(3), 701–718. https://doi.org/10.1007/s11336-019-09675-4.

Liu, Y., & Yang, J. S. (2018). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*, *83*(2), 333–354. https://doi.org/10.1007/s11336-017-9582-9.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245. https://doi.org/10.1007/BF02294018.

Lugannani, R., & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*. https://doi.org/10.1017/s0001867800050278.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*. https://doi.org/10.1177/01466210122031957.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*,. https://doi.org/10.1143/JPSJ.18.328.

Ogasawara, H. (2012). Asymptotic expansions for the ability estimator in item response theory. *Computational Statistics*, *27*(4), 661–683. https://doi.org/10.1007/s00180-011-0282-0.

Ogasawara, H. (2013). Asymptotic properties of the Bayes and pseudo Bayes estimators of ability in item response theory. *Journal of Multivariate Analysis*, *114*, 359–377. https://doi.org/10.1016/J.JMVA.2012.08.013.

Ogasawara, H. (2019). 2019 Asymptotic cumulants of the minimum phi-divergence estimator for categorical data under possible model misspecification. *Communications in Statistics—Theory and Methods*. https://doi.org/10.1080/03610926.2019.1576888.

Rizopoulos, D. (2015). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*(4), 992–1013. https://doi.org/10.1007/s11336-015-9465-x.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342. https://doi.org/10.1007/BF02294437.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627.

Manuscript Received: 18 SEP 2018
Final Version Received: 13 APR 2020
Published Online Date: 16 JUN 2020