PSYCHOMETRIKA—VOL. 83, NO. 2, 333–354
JUNE 2018
https://doi.org/10.1007/s11336-017-9582-9

BOOTSTRAP-CALIBRATED INTERVAL ESTIMATES FOR LATENT VARIABLE SCORES IN ITEM RESPONSE THEORY

YANG LIU AND JI SEUNG YANG

UNIVERSITY OF MARYLAND

In most item response theory applications, model parameters need to be first calibrated from sample data. Latent variable (LV) scores calculated using estimated parameters are thus subject to sampling error inherited from the calibration stage. In this article, we propose a resampling-based method, namely bootstrap calibration (BC), to reduce the impact of the carryover sampling error on the interval estimates of LV scores. BC modifies the quantile of the plug-in posterior, i.e., the posterior distribution of the LV evaluated at the estimated model parameters, to better match the corresponding quantile of the true posterior, i.e., the posterior distribution evaluated at the true model parameters, over repeated sampling of calibration data. Furthermore, to achieve better coverage of the fixed true LV score, we explore the use of BC in conjunction with Jeffreys' prior. We investigate the finite-sample performance of BC via Monte Carlo simulations and apply it to two empirical data examples.

Key words: item response theory, scoring, predictive inference, bootstrap.

1. Introduction

In recent years, advanced item response theory (IRT; e.g., Thissen & Steinberg, 2009) models have gained popularity in not only large-scale assessments but also applied educational, psychological, and health-related research (e.g., Muenks, Wigfield, Yang, & O'Neal, 2017; Curran & Hussong, 2009; Irwin et al., 2010). In IRT, individual differences in the unobserved constructs of interest (e.g., language proficiency, attitude, emotional distress, etc.) are represented by latent variables (LVs). An IRT model has two major components: the distribution of the LV that reflects certain population characteristics of the construct and the response processes through which the observed item responses are related to the LV. The LV score corresponding to an individual response pattern[1] can be estimated from an IRT model, provided all the model parameters that describe the LV distribution and LV-item associations are known.

In practice, LV scores are often calculated in two stages. In the calibration stage, IRT model parameters are estimated from a *calibration sample* by full- or limited-information methods (see Bolt, 2005 for a review); in the current work, we focus on the (marginal) maximum likelihood (ML) estimation (Bock & Lieberman, 1970). In the subsequent scoring stage, response pattern scores are computed with the estimated parameters. The sample of response patterns to be scored, i.e., the *scoring sample*, is not necessarily the same as the calibration sample. In many IRT applications, published model parameter estimates can be borrowed to score the sample at hand, without any need for recalibration (e.g., Magnus et al., 2016). The two-stage estimates of LV scores are subject to two sources of uncertainty: *measurement error* and *sampling error*. Measurement error refers to the fact that test items are imperfect indicators of the LVs; it decreases as the length of the test tends to infinity, as a result of the increasing reliability of the measurement scale (e.g.,

Correspondence should be made to Yang Liu, Department of Human Development and Quantitative Methodology, University of Maryland, 1230B Benjamin Building, College Park, MD 20742 USA. Email: yliu87@umd.edu

[1] We only consider response pattern scoring here; however, our discussions can be extended to IRT scoring based on summed scores.

Sireci, Thissen, & Wainer, 2001). Sampling error is carried over from the use of estimated model parameters that differ from their true values; it becomes negligible when the calibration sample is sufficiently large.

Conventionally, estimated IRT model parameters are used in calculating LV scores as though the true parameters were known, leading to the so-called *plug-in* method. The plug-in method ignores sampling error and thus often yields overstated measurement precision when the calibration sample size is limited (e.g., Cheng & Yuan, 2010; Patton, Cheng, Yuan, & Diao, 2014; Mislevy, Wingersky & Sheehan, 1993; Yang, Hansen, & Cai, 2012). It may also incur biased inferences in the further use of scores, e.g., computerized adaptive testing (Patton, Cheng, Yuan, & Diao, 2013). To account for the influence of sampling error, existing methods in the literature mostly focus on upward-adjusting the standard errors (SEs) associated with the LV scores which originally reflect only measurement error. For example, Cheng & Yuan (2010) derived an SE correction for the ML scores using a Taylor series expansion argument. Yang et al. (2012) proposed a multiple-imputation (MI)-based characterization for sampling and measurement error in the context of Bayesian scoring (i.e., based on the posterior distribution of the LV conditional on the response pattern to be scored; see Sect. 2.1), which extends the work of Mislevy et al. (1993).

As a standard tool for uncertainty characterization, interval estimation[2] provides a range of possible values for an inferential target, i.e., a parameter or random variable, so as to facilitate a probabilistic interpretation of the result. Proper interval estimators for LV scores should reflect both sources of uncertainty. Under the Bayesian scoring framework (see Sect. 2.1 for more details), measurement error is typically gauged by the variability of the LV posterior distribution evaluated at the true IRT model parameters, while sampling error is reflected by the discrepancy between the estimated and true posterior distributions. In line with the classic empirical Bayes examples discussed by Barndorff-Nielsen and Cox (1996), we investigate the interval estimation of LV scores from the perspective of predictive inference: Plausible values generated from the true posterior are treated as the "future" data, which we would like to predict using IRT model parameters estimated from the "current" calibration data. An interval estimator of the LV score is expected to cover the true posterior plausible values with some prescribed probability over repeated sampling of both the plausible values and calibration data. The difference between actual and intended coverage in the prediction problem is a natural quantification of sampling error.

Motivated by the literature of predictive inference (Beran, 1990; Fonseca, Giummolè, & Vidoni, 2014), we propose a resampling-based modification to the plug-in interval estimates of LV scores, termed bootstrap calibration (BC). The BC-based interval estimator typically attains the nominal predictive coverage level at a faster rate compared to the corresponding plug-in one. The finite-sample performance of BC-based and plug-in interval estimators is evaluated via Monte Carlo simulations. Then, we consider combining BC and Bayesian scoring under Jeffreys' prior to improve the coverage of the true LV score—the fixed LV value that generates the response pattern being scored. Two empirical data examples are presented in the end.

[2] The intervals discussed in the current work may be labeled confidence intervals, credible intervals, or prediction intervals depending on the context and one's philosophy toward statistical inference. For simplicity, we use the unified name "interval estimate/estimator." The interpretation of an interval estimate should be determined by the inferential target and the definition of coverage probability.

2. Bootstrap Calibration: Theory and Method

2.1. A Bayesian Framework for Scoring

Let θ_i be a unidimensional LV for person i: θ_i is continuous with density $\phi(\cdot; \boldsymbol{\gamma})$ supported on the real line $\mathbb{R}$, in which $\boldsymbol{\gamma}$ denotes the unknown parameters involved in the density function. Many IRT software packages assume by default that θ_i follows $\mathcal{N}(0, 1)$, in which no parameters need to be estimated. Other families of LV distributions (e.g., Woods & Thissen, 2006; Noel & Dauvier, 2007) may include parameters to be estimated from the data. Let Y_{ij} be person i's response to a K-category item j. Conditional on $\theta_i = \theta$, the probability of $Y_{ij} = k, k \in \{0, \ldots, K-1\}$, is given by the item response function (IRF) $f_j(k|\theta; \boldsymbol{\beta}_j)$, in which $\boldsymbol{\beta}_j$ denotes the item parameters. For dichotomous response data, the one-, two-, and three-parameter logistic (1–3PL) IRFs are typically used (Birnbaum, 1968). For ordinal response data, the IRFs are often formulated using the cumulative or adjacent logit link functions, leading to the graded response model (Samejima, 1969) or the generalized partial credit model (Muraki, 1992).

Now consider a test of m items. Assume the item responses $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\top$ are independent given the latent variable θ_i (*local independence*; McDonald, 1981): i.e.,

$$f(\mathbf{y}_i|\theta; \boldsymbol{\beta}) = \prod_{j=1}^{m} f_j(y_{ij}|\theta; \boldsymbol{\beta}_j), \tag{1}$$

in which $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top$ denotes a fixed response pattern, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$. The left-hand side of Eq. 1 gives the *conditional* probability of $\mathbf{Y}_i = \mathbf{y}_i$ given $\theta_i = \theta$. Then, the *joint* density function of $\mathbf{Y}_i = \mathbf{y}_i$ and $\theta_i = \theta$ can be expressed as

$$f(\mathbf{y}_i, \theta; \boldsymbol{\xi}) = f(\mathbf{y}_i|\theta; \boldsymbol{\beta}) \cdot \phi(\theta; \boldsymbol{\gamma}), \tag{2}$$

in which $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ collects all the item and LV-density parameters.

If $\boldsymbol{\xi}$ is known, Bayesian inferences about LV scores can be made from the *posterior* distribution of θ_i given response pattern $\mathbf{y}_i$, which is determined by the following density function:

$$g(\theta|\mathbf{y}_i; \boldsymbol{\xi}) = \frac{f(\mathbf{y}_i, \theta; \boldsymbol{\xi})}{f(\mathbf{y}_i; \boldsymbol{\xi})} \propto f(\mathbf{y}_i, \theta; \boldsymbol{\xi}). \tag{3}$$

in which the denominator is the *marginal* probability[3] of $\mathbf{Y}_i = \mathbf{y}_i$, denoted

$$f(\mathbf{y}_i; \boldsymbol{\xi}) = \int_{-\infty}^{\infty} f(\mathbf{y}_i, \theta; \boldsymbol{\xi}) \mathrm{d}\theta. \tag{4}$$

The population distribution of the LV with density $\phi(\cdot; \boldsymbol{\gamma})$ functions as a *prior* distribution in the posterior calculation; accordingly, both names are used interchangeably in the sequel. The *mean* of the posterior distribution, i.e., $\bar{\theta}(\mathbf{y}_i; \boldsymbol{\xi}) = \int_{-\infty}^{\infty} \theta g(\theta|\mathbf{y}_i; \boldsymbol{\xi}) \mathrm{d}\theta$, gives the expected a posteriori (EAP) score (Thissen & Wainer, 2001, p. 112; Bock & Mislevy, 1982; Lazarsfeld, 1950, p. 464), and the posterior standard deviation (SD) $\sigma(\mathbf{y}_i; \boldsymbol{\xi}) = \sqrt{\int_{-\infty}^{\infty} [\theta - \bar{\theta}(\mathbf{y}_i; \boldsymbol{\xi})]^2 g(\theta|\mathbf{y}_i; \boldsymbol{\xi}) \mathrm{d}\theta}$ quantifies (the lack of) measure precision. A $100\alpha\%$ normal approximation interval estimator for the

[3] When Eq. 4 is viewed as a function of $\boldsymbol{\xi}$, it is typically referred to as the marginal likelihood.

LV score is given by $\bar{\theta}(\mathbf{y}_i; \boldsymbol{\xi}) \pm z_{(1+\alpha)/2}\sigma(\mathbf{y}_i; \boldsymbol{\xi})$, in which $z_{(1+\alpha)/2}$ stands for the $[(1+\alpha)/2]$th quantile of the standard normal distribution.[4]

Alternatively, interval estimates of scale scores can be constructed from the quantiles of the posterior distribution. Let

$$G(\theta|\mathbf{y}_i; \boldsymbol{\xi}) = \int_{-\infty}^{\theta} g(\theta'|\mathbf{y}_i; \boldsymbol{\xi})\mathrm{d}\theta' = \frac{\int_{-\infty}^{\theta} f(\mathbf{y}_i, \theta'; \boldsymbol{\xi})\mathrm{d}\theta'}{\int_{-\infty}^{\infty} f(\mathbf{y}_i, \theta; \boldsymbol{\xi})\mathrm{d}\theta} \tag{5}$$

be the posterior cumulative distribution function (cdf). Fix $\boldsymbol{\xi}$ and $\mathbf{y}_i$ for now. Provided the cdf $G(\cdot|\mathbf{y}_i; \boldsymbol{\xi})$ is continuous and strictly increasing, the quantile function $G^{-1}(\cdot|\mathbf{y}_i; \boldsymbol{\xi})$ is the unique inverse of the cdf that satisfies

$$G(G^{-1}(\alpha|\mathbf{y}_i; \boldsymbol{\xi})|\mathbf{y}_i; \boldsymbol{\xi}) = \alpha \tag{6}$$

for all $\alpha \in (0, 1)$. For commonly used IRT models with normal LV densities and logit/probit link functions (e.g., the 3PL model, the graded response model, etc.), the joint density $f(\mathbf{y}_i, \theta; \boldsymbol{\xi})$ is often nonzero across the domain of θ, which guarantees the continuity and strict monotonicity of the posterior cdf.

2.2. *Characterizing Sampling Error: A Prediction Problem*

From now on, write $\boldsymbol{\xi}_0$ as the *true population values* of the IRT model parameters. Since $\boldsymbol{\xi}_0$ is unknown in reality, it needs to be estimated from a calibration sample. Denote by $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\top$ a calibration sample composed of n independent and identically distributed (i.i.d.) response patterns, each of which follows the probability mass function $f(\cdot; \boldsymbol{\xi}_0)$. Under necessary regularity conditions (Birch, 1964; Bishop, Fienberg, & Holland, 1975, Chapter 14.8) the ML estimator of model parameters, denoted $\hat{\boldsymbol{\xi}}$, is asymptotically normal and efficient:

$$\sqrt{n}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}(\boldsymbol{\xi}_0)^{-1}\right). \tag{7}$$

In Eq. 7, $\mathcal{I}(\boldsymbol{\xi})$ denotes the Fisher information matrix:

$$\mathcal{I}(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}}^{\mathbf{Y}_i}\left[\frac{\partial \log f(\mathbf{Y}_i; \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \frac{\partial \log f(\mathbf{Y}_i; \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^\top}\right] = -E_{\boldsymbol{\xi}}^{\mathbf{Y}_i}\left[\frac{\partial^2 \log f(\mathbf{Y}_i; \boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^\top}\right], \tag{8}$$

in which the expectation $E_{\boldsymbol{\xi}}^{\mathbf{Y}_i}$ is taken over repeated sampling of $\mathbf{Y}_i$ from the correctly specified IRT model with parameters $\boldsymbol{\xi}$. It can also be established under stronger assumptions that the bias of the ML estimator is of order n^{-1} (Cox & Snell, 1968; see also Appendix A.1).

Among standard regularity conditions, the global identification condition (Birch, 1964, Condition B) remains an open problem for general IRT models. Relevant discussions can be found in San Martin, Rolin, and Castro (2013) and San Martin (2016). To acknowledge the importance of model identification, we examine in our simulation study (see Sect. 3) the local identification of the data generating model: i.e., the positive definiteness of the Fisher information matrix when evaluated at the true item parameters. Local identification also suffices for the theoretical justification of the proposed method.

We denote the response pattern to be scored by $\mathbf{y}_0$, in order to distinguish it from the calibration sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. For conciseness, the conditioning on $\mathbf{y}_0$ is suppressed in the notations of the posterior density, cdf, and quantiles (Eqs. 3, 5, 6). The sampling error inherent in LV scores can be

[4] For notational succinctness, we use α for coverage in the current work; by convention, however, $1-\alpha$ is typically used. As a result, the $[(1+\alpha)/2]$th quantile in our notation is the same as the $(1-\alpha/2)$th quantile in the conventional notation.

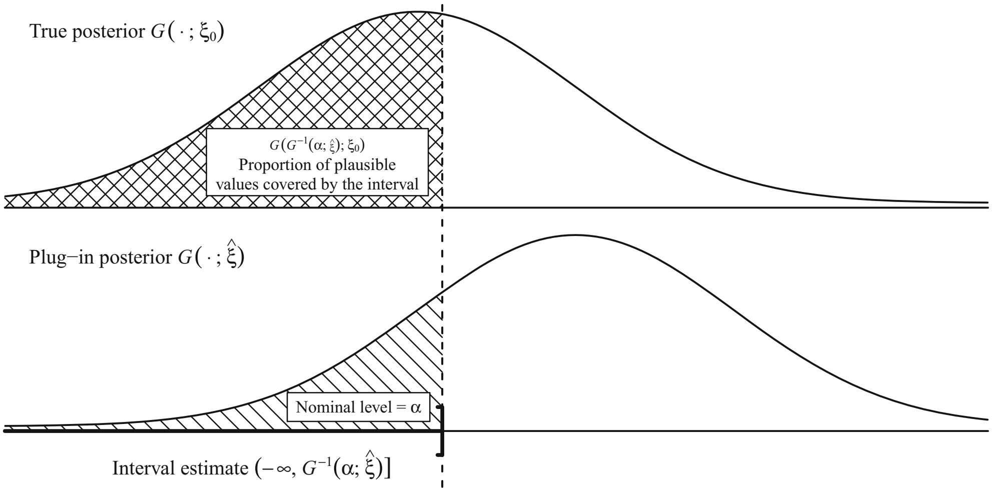
True posterior $G(\cdot\,;\xi_0)$
$G(G^{-1}(\alpha;\hat{\xi});\xi_0)$
Proportion of plausible values covered by the interval
Plug–in posterior $G(\cdot\,;\hat{\xi})$
Nominal level = α
Interval estimate $(-\infty, G^{-1}(\alpha;\hat{\xi})]$

FIGURE 1.
The discrepancy between the true and plug-in posteriors and an interpretation pertaining to predictive inference. The αth quantile of the plug-in posterior is denoted by $G^{-1}(\alpha;\hat{\boldsymbol{\xi}})$, which defines a one-sided interval estimate with nominal level α (*outlined in bold*). Evaluating the true cdf at the plug-in quantile gives $G(G^{-1}(\alpha;\hat{\boldsymbol{\xi}});\boldsymbol{\xi}_0)$, i.e., the *crosshatched* area under the true posterior, which in general is not equal to α. The *crosshatched* area also gives the proportion of plausible values (generated from the true posterior) covered by the interval estimate.

intuitively construed as the discrepancy between the true posterior $G(\cdot;\boldsymbol{\xi}_0)$ and some estimated version derived from $\hat{\boldsymbol{\xi}}$. The plug-in posterior $G(\cdot;\hat{\boldsymbol{\xi}})$ is the simplest and most commonly used estimate of $G(\cdot;\boldsymbol{\xi}_0)$. Due to the consistency of $\hat{\boldsymbol{\xi}}$, $G(\theta;\hat{\boldsymbol{\xi}})$ approaches $G(\theta;\boldsymbol{\xi}_0)$ for each θ as the calibration sample size n tends to infinity. One way to quantify the difference between the plug-in and true posterior distributions is to calculate the following quantity

$$C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0) = G(G^{-1}(\alpha;\hat{\boldsymbol{\xi}});\boldsymbol{\xi}_0) \tag{9}$$

and to compare it with α. If the true posterior is well approximated by the plug-in estimate, then Eq. 9 should yield approximately α for all $\alpha \in (0, 1)$.

The discrepancy measure defined in Eq. 9 has an alternative interpretation pertaining to predictive inference. Let Q be a random variable that follows the true posterior $G(\cdot;\boldsymbol{\xi}_0)$; realizations of Q are often called *plausible values* of LV scores from the true posterior. By definition, the cdf $G(\theta;\boldsymbol{\xi}_0)$ designates the chance for Q to fall at or below θ. Consequently, $C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0)$ amounts to the probability that Q is covered by the one-sided plug-in interval estimate $(-\infty, G^{-1}(\alpha;\hat{\boldsymbol{\xi}})]$, conditional on the ML estimates $\hat{\boldsymbol{\xi}}$. We say the *nominal level* of such an interval estimate is α, in view of the fact that the αth plug-in posterior quantile is used as a proxy of the corresponding true posterior quantile which would have resulted in exact coverage probability α. This new perspective on the characterization of sampling error is illustrated in Fig. 1.

The *predictive coverage* of the plug-in interval estimator $(-\infty, G^{-1}(\alpha;\hat{\boldsymbol{\xi}})]$ is defined as the average of $C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0)$ over repeated samples of the calibration data $\mathbf{Y}$, i.e.,

$$C(\alpha;\boldsymbol{\xi}_0) = E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}\left[C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0)\right] = E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}\left[G(G^{-1}(\alpha;\hat{\boldsymbol{\xi}});\boldsymbol{\xi}_0)\right], \tag{10}$$

in which $E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}$ denotes the expectation over the data space under the true model. In the literature of predictive inference (e.g., Beran, 1990; Barndorff-Nielsen & Cox, 1996), predictive coverage serves as an important criterion to evaluate prediction intervals. Next, we discuss the asymptotic properties of the plug-in interval estimator and a resampling-based procedure to improve its predictive coverage.

2.3. *Predictive Calibration*

Although the predictive coverage of the plug-in interval estimator (Eq. 10) converges to its nominal level α as $n \to \infty$, the deviation can be substantial when n is small. Because the bias of the ML estimator is typically of order n^{-1}, it is conceivable that the difference between $C(\alpha; \boldsymbol{\xi}_0)$ and α is also of order n^{-1}. The result can be established by a Taylor series expansion argument, which has been discussed by Beran (1990) and Barndorff-Nielsen and Cox (1996). More details can be found in Appendix A.2.

An adjustment to the one-sided plug-in interval, which was originally proposed by Beran (1990), can be applied to reduce the $O(n^{-1})$ remainder term of the predictive coverage to $O(n^{-2})$; Beran named such a method "calibration." To avoid confusion with the calibration of test items (i.e., estimating IRT model parameters), we refer to this adjustment as *predictive calibration*. Let $C^{-1}(\cdot; \boldsymbol{\xi}_0)$ be the solution of

$$C(C^{-1}(\alpha; \boldsymbol{\xi}_0); \boldsymbol{\xi}_0) = \alpha, \tag{11}$$

which often uniquely exists due to the continuity and strict monotonicity of $G(\cdot; \boldsymbol{\xi})$ for all $\boldsymbol{\xi}$ in the closure of the parameter space.[5] Equation 11 suggests that the predictive coverage of a one-sided plug-in interval with nominal level $C^{-1}(\alpha; \boldsymbol{\xi}_0)$ is exactly α. Utilizing this fact to achieve exact predictive coverage, although theoretically appealing, is not viable in practice because the true IRT model parameters $\boldsymbol{\xi}_0$ remain unknown. Nevertheless, the approximation to the desired predictive coverage α can still be improved by modifying the nominal level from α to the "best guess" of $C^{-1}(\alpha; \boldsymbol{\xi}_0)$, i.e., $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$, which yields the method of predictive calibration. The predictive coverage of the resulting interval estimator $(-\infty, G^{-1}(C^{-1}(\alpha; \hat{\boldsymbol{\xi}}); \hat{\boldsymbol{\xi}})]$

$$\tilde{C}(\alpha; \boldsymbol{\xi}_0) = E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}\left[G(G^{-1}(C^{-1}(\alpha; \hat{\boldsymbol{\xi}}); \hat{\boldsymbol{\xi}}); \boldsymbol{\xi}_0)\right] \tag{12}$$

often equals to α plus a residual term of order $O(n^{-2})$. A proof in the context of general predictive inference can be found in Beran (1990); we outlined in "Appendix A.3" a simplified argument that is specialized for the scoring problem. The required assumptions can be verified for most commonly used IRT models such as the 3PL model and the graded response model.

Alternatively, asymptotic matching of the predictive coverage can be accomplished by direct analytic calculations (Barndorff-Nielsen & Cox, 1996; Beran, 1990; Fonseca et al., 2014; Vidoni, 1998; 2009). Those analytic corrections require calculating the second-order partial derivatives of the posterior cdf (Eq. 5) and the Fisher information matrix (Eq. 8) with respect to $\boldsymbol{\xi}$, which can be computationally challenging for IRT models. In the meantime, interval estimates adjusted by predictive calibration can be obtained via parametric bootstrap (see Sect. 2.4), which does not entail explicit calculations of the derivatives and can be easily implemented for a wide variety of IRT models using existing software packages.

[5] When the parameter space is unbounded, the ML estimates can be infinite and the corresponding posterior cdf may have jumps. Those irregular cases often happen with only exponentially small probability and can be removed from the calculation of $C(\cdot; \boldsymbol{\xi}_0)$ and $C^{-1}(\cdot; \boldsymbol{\xi}_0)$.

Beran (1990) commented that predictive calibration can be iterated to obtain higher-order matching results under stronger regularity conditions. In other words, if we apply a similar predictive calibration argument to the calibrated predictive coverage $\tilde{C}(\alpha; \boldsymbol{\xi}_0)$ (Eq. 12), then typically the approximation error can be further reduced to $O(n^{-3})$. As described in Sect. 2.4, predictive calibration is often carried out via parametric bootstrap; therefore, one more iteration of predictive calibration introduces nested resampling, which leads to a remarkable increase in the computation load and thus reduces its utility. In addition, we also found in our simulation study (Sect. 3) that one round of predictive calibration is often enough to achieve accurate predictive coverage at most α levels, even when the calibration sample size is small ($n = 250$).

2.4. *Computation via Parametric Bootstrap*

Our remaining task is to calculate $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$, the adjusted nominal level in predictive calibration. The method we develop is an application of Beran's (1990) Algorithm 1.

Conditional on the ML estimates $\hat{\boldsymbol{\xi}}$, the function $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$ we intend to approximate is the inverse function of $C(\alpha; \hat{\boldsymbol{\xi}})$, which can be expanded as:

$$C(\alpha; \hat{\boldsymbol{\xi}}) = E_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star}\left[G(G^{-1}(\alpha; \hat{\boldsymbol{\xi}}^\star); \hat{\boldsymbol{\xi}})\right]. \tag{13}$$

In Eq. 13, $E_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star}$ denotes the expectation over resampling of i.i.d. calibration data $\mathbf{Y}^\star$ from the IRT model with parameters $\hat{\boldsymbol{\xi}}$ (which is considered fixed here), and $\hat{\boldsymbol{\xi}}^\star$ denotes the ML estimates obtained from $\mathbf{Y}^\star$. The generation of $\mathbf{Y}^\star$ and the calculation of $\hat{\boldsymbol{\xi}}^\star$ coincide with the procedure of parametric bootstrap (e.g., Efron & Tibshirani, 1994); hence, we term our procedure bootstrap calibration, abbreviated as BC. It is emphasized that we use parametric bootstrap to approximate the expectation $E_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star}$ involved in the adjusted nominal level, which differs in essence from, e.g., Patton et al.'s (2014) use of bootstrap to calculate the SEs of LV scores.

Although it is possible to calculate the inverse function of Eq. 13 via numerical root finding algorithms such as Brent's method (Brent, 1973), the computational cost of function evaluations is often too high. In Beran's algorithm, $C(\alpha; \hat{\boldsymbol{\xi}})$ is re-expressed as a cdf which can be approximated by the corresponding empirical cdf of a Monte Carlo sample; as a result, $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$ can be treated as a quantile function, i.e., the inverse of a cdf, which can be efficiently approximated by interpolating empirical quantiles.

Conditional on $\hat{\boldsymbol{\xi}}$, let $Q^\star$ be a random variable that follows the plug-in posterior $G(\cdot; \hat{\boldsymbol{\xi}})$. The right-hand side of Eq. 13 can be further written as

$$E_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star}\left[G(G^{-1}(\alpha; \hat{\boldsymbol{\xi}}^\star); \hat{\boldsymbol{\xi}})\right] = P_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star, Q^\star}\left\{Q^\star \le G^{-1}(\alpha; \hat{\boldsymbol{\xi}}^\star)\right\} = P_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star, Q^\star}\left\{G(Q^\star; \hat{\boldsymbol{\xi}}^\star) \le \alpha\right\}, \tag{14}$$

in which $P_{\hat{\boldsymbol{\xi}}}^{\mathbf{Y}^\star, Q^\star}\{\cdot\}$ denotes the joint probability measure of $\mathbf{Y}^\star$ and $Q^\star$. In Eq. 14, the first equality follows from a graphical illustration resembling Fig. 1, in which the plug-in posterior $G(\cdot; \hat{\boldsymbol{\xi}})$ is replaced by the bootstrap posterior $G(\cdot; \hat{\boldsymbol{\xi}}^\star)$, and the true posterior $G(\cdot; \boldsymbol{\xi}_0)$ by the plug-in posterior. The second equality is obtained by applying $G(\cdot; \hat{\boldsymbol{\xi}}^\star)$ on both sides of the inequality inside the bracket. The right-hand side of Eq. 14 corresponds to the cdf of the random variable $G(Q^\star; \hat{\boldsymbol{\xi}}^\star)$ evaluated at α. Therefore, $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$ amounts to the αth quantile of $G(Q^\star; \hat{\boldsymbol{\xi}}^\star)$, which can be further approximated by Monte Carlo sampling. For easy reference, we summarize the computation of BC-based one-sided interval estimates for response pattern LV scores as Algorithm 1.

Algorithm 1 Bootstrap calibration

```
1: Input: $\hat{\boldsymbol{\xi}}$—the ML estimates, $\mathbf{y}_0$—the response pattern to be scored
2: Generate plausible values $Q^\star = q^{\star(b)}$, $b = 1, \ldots, B$, from the plug-in posterior $G(\cdot; \hat{\boldsymbol{\xi}})$
3: for all $b = 1, \ldots, B$ do
4:     Generate a bootstrap sample $\mathbf{y}^{\star(b)}$ using estimated parameters $\hat{\boldsymbol{\xi}}$
5:     Obtain the ML estimates $\hat{\boldsymbol{\xi}}^{\star(b)}$ from the bootstrap sample $\mathbf{y}^{\star(b)}$
6:     Calculate $G(q^{\star(b)}; \hat{\boldsymbol{\xi}}^{\star(b)})$
7: end for
8: Approximate $C^{-1}(\alpha; \hat{\boldsymbol{\xi}})$ by the empirical $\alpha$th quantile of $G(q^{\star(b)}; \hat{\boldsymbol{\xi}}^{\star(b)})$, $b = 1, \ldots, B$
```

There is more than one way to simulate plausible values of LV scores from the fitted IRT model (Line 2 of Algorithm 1). In our simulation studies (Sects. 3, 5) and empirical example (Sect. 6), we implement a random-walk Metropolis (Metropolis et al., 1953) algorithm that has been routinely used for plausible value generation in existing IRT software packages, such as IRTPRO (Cai, Thissen, & du Toit, 2011), flexMIRT (Houts & Cai, 2013), and `mirt` (Chalmers, 2012). The random walk is generated by a zero-mean Gaussian incremental variate whose variance parameter is selected based on pilot runs such that the acceptance ratio falls between 0.3 and 0.4.

Line 6 of Algorithm 1 also calls for the evaluation of the posterior cdf (Eq. 5) at the bootstrap ML estimates $\hat{\boldsymbol{\xi}}^{\star(b)}$, which involves two integrations of the joint density function (Eq. 2). To ensure an accurate approximation to the ratio of improper integrals in the posterior cdf, we suggest the use of rescaled Gauss–Legendre quadrature with properly chosen end points. When a standard normal distribution is assumed for the LV, the doubly improper integral in the denominator of Eq. 5, i.e., the marginal probability of a response pattern, is typically approximated by Gauss–Hermite or rectangular quadrature. We found in our simulation work that a rescaled 61-point Gauss–Legendre quadrature (originally defined on $[-1, 1]$) defined on $[-6, 6]$ and a gold-standard 61-point Gauss–Hermite quadrature yield almost identical approximations to the marginal probability. In addition, it is noted that the posterior cdf can also be approximated empirically by generating a Monte Carlo sample from the posterior distribution, i.e., Beran's Algorithm 2 or double bootstrap; however, the double bootstrap is computationally more demanding than the direct numerical integration for unidimensional IRT models and hence is not considered further.

3. Simulation 1

In this section, we report a simulation study in which BC-based and plug-in interval estimators of LV scores were compared in terms of predictive coverage (i.e., Eqs. 10, 12). A test of $m = 36$ items was considered, and the true data generating model was 3PL with a constant pseudo-guessing parameter. The true discrimination and difficulty parameters were randomly sampled from $\mathcal{U}(0.5, 2)$ and $\mathcal{U}(-2, 2)$, respectively, and the pseudo-guessing parameter was set to 0.2. The range of item parameter values was specified following Rupp (2013); Fig. 2 displays histograms of the randomly generated item discriminations and difficulties. To verify the local identifiability of the data generating model, we calculated the observed information matrix using 500,000 simulated response patterns.[6] The eigenvalues of the observed information matrix range from 0.003 to 0.521, which suggests that the true parameter values form a local maximum of the likelihood function.

[6] A direct calculation of the expected information is not viable because there are $2^{36} \approx 6.87 \times 10^{10}$ response patterns. The observed information is a consistent estimator of the expected information.

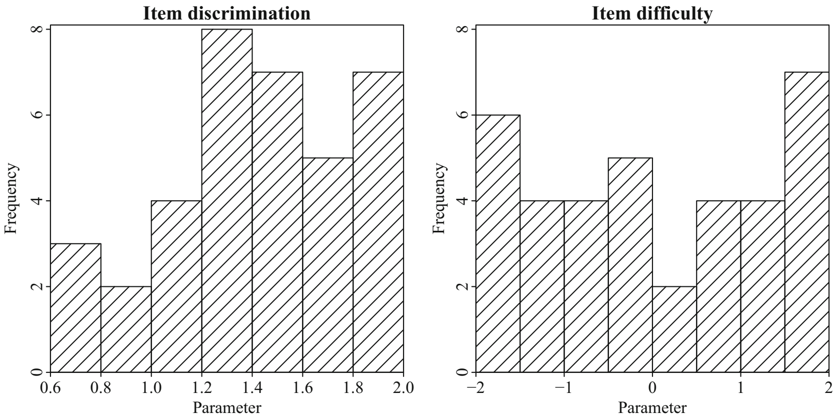

FIGURE 2.
Histograms of item discrimination (*left panel*) and difficulty (*right panel*) parameters in the data generating model.

The fitted model was 3PL with the pseudo-guessing parameters constrained equal across items.[7] Two calibration sample size conditions were considered: $n = 250$ and 500; 500 replications were completed under each condition. In each replication, we generated five response patterns from the true 3PL model at the following equally spaced LV score levels: $\theta_0 = -2, -1$, 0, 1, and 2. For each pattern to be scored, one-sided plug-in and BC-based interval estimates were constructed at 39 equally spaced nominal α levels: $\alpha = 0.025, 0.05, \ldots, 0.975$.

The software package `mirt` (Chalmers, 2012) in the statistical computing environment R (R Core Team, 2016) was used for data generation and model fitting. `mirt` implements an expectation-maximization (EM) algorithm (Bock and Aitkin, 1981) to find the ML estimates of 3PL item parameters. We adopted the software's default settings of numerical quadrature (61 equally spaced rectangular quadrature points from −6 to 6), convergence criterion (maximum absolute change of parameters <0.0001), and allowed maximum number of iterations (500). To perform BC, 500 bootstrap samples were generated in each replication. Occasionally, model fittings in bootstrap samples may fail to coverage within 500 iterations; if that happened, we cycled the loop and generated a new bootstrap sample. A rescaled 61-point Gauss–Legendre quadrature from −6 to 6 was used to approximate the posterior cdf and quantile function in both BC and the calculation of predictive coverage. The predictive coverage of one-sided interval estimators, i.e., Eq. 10 for the plug-in method and Eq. 12 for BC, requires the calculation of posterior quantiles. The R function `uniroot`, which implements a modified version of Brent's (1973) algorithm, was used to numerically find the solution of Eq. 6.

Figure 3 displays the average deviation between the empirical predictive coverage and the nominal α level pooling across 500 repeated samples of the calibration and scoring data; the results were summarized by the true LV score value θ_0. In small calibration samples ($n = 250$), the coverage of the plug-in interval can be substantially off when $\theta_0 \geq 1$; at $\theta_0 = 2$, the difference is more than 0.08 when α is between 0.6 and 0.9. The BC-based interval, in contrast, exhibits more accurate empirical coverage in small samples: The largest deviation from the nominal level is 0.02 when $\theta_0 = 2$ and $\alpha \in [0.55, 0.6]$. As the sample size increases to $n = 500$, the performance of the plug-in interval improves, consistent with the asymptotic theory discussed in Sect. 2.3. However, its empirical predictive coverage is still lower than the nominal level by more than 0.025 at the

[7] Fitting the full-rank 3PL model with unconstrained pseudo-guessing parameters in small samples proves to be challenging (e.g., Han, 2012). The constrained version, however, seldom caused any convergence issue in our simulation study even when the sample size is only 250.

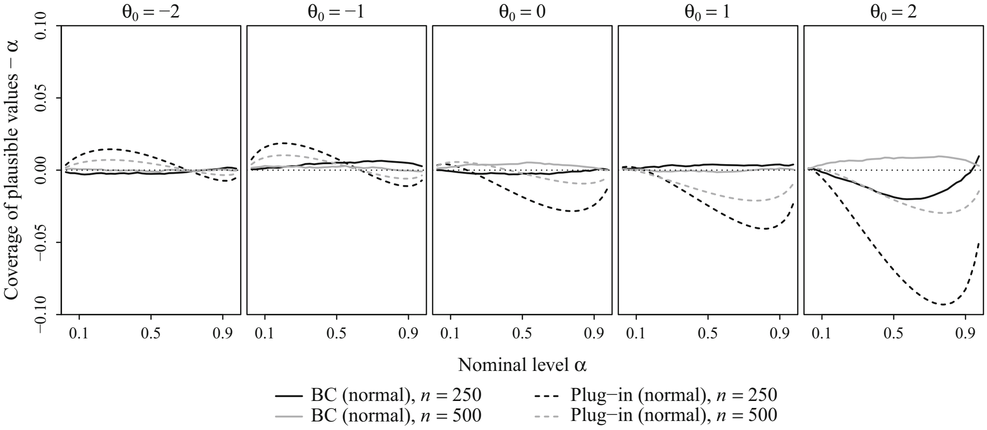

FIGURE 3.
Empirical predictive coverage for one-sided interval estimates of LV scores (sample size $n = 250$ and 500, test length $m = 36$, normal prior). Simulation results for different true LV score levels ($\theta_0 = -2, -1, 0, 1$, and 2) are displayed in separate panels. In *each panel*, the x-axis is the nominal α level and the y-axis is the discrepancy between the empirical predictive coverage and the nominal α level. Results under the two sample size conditions are depicted in different colors: *black* for $n = 250$ and *gray* for $n = 500$. *Solid* and *dashed lines* represent the empirical coverage values for bootstrap calibration (BC) and plug-in methods, respectively.

most extreme LV score level ($\theta_0 = 2$) for α between 0.6 and 0.9. After predictive calibration, the difference is reduced to no more than 0.01 across all the θ_0 and α levels.

Per our discussion in Sect. 2.2, predictive coverage gauges the extent to which sampling error in estimated IRT model parameters is properly captured by the interval estimators of LV scores. Our simulation results suggest that the BC-based interval estimator outperforms the plug-in one in terms of predictive coverage at most true LV score and nominal α levels, especially when the sample size is only 250. BC is thus recommended for response pattern scoring with small calibration samples. In large samples ($n = 500$), the advantage of BC, albeit noticeable at certain θ_0 levels, is in general not major. We also noted that two-sided quantile-based interval estimates are more frequently used in practice than one-sided ones; for example, the interval between the 0.05th and 0.95th quantiles of the plug-in posterior $G(\cdot; \hat{\boldsymbol{\xi}})$ gives an equi-tailed plug-in interval estimate of LV scores at nominal level $\alpha = 0.9$. When two-sided interval estimation is desired, BC can be applied to both limits of the interval. The resulting two-sided intervals should also be preferred over the plug-in intervals owing to the superiority of BC across most α levels in the one-sided case.

4. Bayesian Scoring with Jeffreys' Prior

4.1. Coverage of the True Latent Variable Score

So far, we have focused on the predictive coverage of interval estimators, which is a natural discrepancy measure between the true and estimated posterior distributions of the LV conditional on observed item responses (Eq. 3 or 5). In the Bayesian scoring framework described in Sect. 2.1, the LV is treated as a *random effect* in the population; the posterior distribution quantifies the plausibility of a person's LV score given the observed response pattern, assuming that the person is sampled randomly from the population. A competing convention of IRT scoring, which is popular in educational measurement, treats each individual score as a *fixed effect*. This fixed-effect approach is the basis of, for example, ML scoring (Thissen & Wainer, 2001, pp. 100–103;

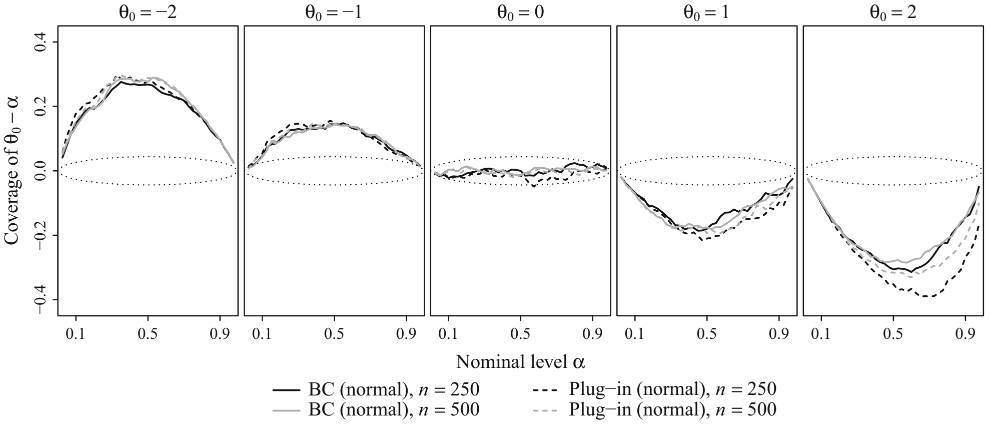

FIGURE 4.
Empirical coverage of θ_0 for one-sided interval estimates of LV scores (sample size $n = 250$ and 500, test length $m = 36$, normal prior). Simulation results for different true LV score levels ($\theta_0 = -2, -1$, 0, 1, and 2) are displayed in separate panels. In *each panel*, the x-axis is the nominal α level, and the y-axis is the discrepancy between the empirical coverage of θ_0 and the nominal α level. Results under the two sample size conditions are depicted in different colors: *black* for $n = 250$ and *gray* for $n = 500$. *Solid* and *dashed lines* represent the empirical coverage values for bootstrap calibration (BC) and plug-in methods, respectively. The *dotted curves* delineate a 95% pointwise confidence envelope for the coverage probability based on normal approximation.

Lazarsfeld, 1950, p. 464) and joint ML (JML) estimation (Baker & Kim, 2004, pp. 84–107; Birnbaum, 1968, p. 420). A referee suggested San Martin and De Boeck (2015) and San Martin (2016) as further readings on random- versus fixed-effect IRT models. Next, we discuss the interval estimation of LV scores under the fixed-effect scoring framework: the response pattern being scored, which has been denoted by $\mathbf{y}_0$, is assumed to be generated from some fixed true score θ_0 through the correctly specified conditional probability (Eq. 1). One clarification to make is that we only adopt the fixed-effect approach in the scoring stage, while we still calibrate the measurement model via the (marginal) ML in which a population distribution of the LV is assumed.

When the fixed-effect approach is embraced, the inferential target becomes the true score θ_0, in lieu of plausible values generated from the true posterior. Bayesian methods based on the posterior distribution of LV scores, regardless of the choice of priors, yield asymptotically correct inferences for θ_0 as the number of test items m tends to infinity, which is a direct consequence of the celebrated Bernstein–von Mises Theorem (e.g., Le Cam & Yang, 2000; see also Chang & Stout, 1993). However, given that many measurement scales in educational and psychological research (e.g., emotional distress measures) are composed of no more than a handful of items, the shrinkage of LV score estimates incurred by the prior distribution of LV is often substantial, especially when informative priors, such as a standard normal distribution, are assumed. Although the shrinkage is desirable when certain Bayesian optimality is intended, it could have an adverse impact on the coverage of θ_0 in the frequentist sense.

To illustrate the effect of shrinkage, we revisited the data generated in our previous simulation study (Sect. 3) and plotted the *coverage of true* θ_0 for the two types of one-sided interval estimators at various α and θ_0 levels (see Fig. 4). It is observed that the discrepancy between the empirical coverage of θ_0 and the nominal α level increases as θ_0 deviates from the prior mean of 0; the maximum discrepancy is about 0.4 when $\theta_0 = -2$ and α is between 0.4 and 0.5. The BC-based interval performs better than the plug-in interval; however, the improvement is quite trivial. It is emphasized that the inadequacy in the coverage of θ_0 results from the limited test length and the use of informative priors, which cannot be mitigated by increasing the calibration sample size.

In fact, we found by comparing the black and gray lines in Fig. 4 that increasing the calibration sample size only makes the two types of interval estimators behave more similarly.

When a practical decision concerns individual scores and the associated measurement precision, it is important to recognize the influence of prior selection. Next, we study the degree to which weakly informative Bayesian scoring using Jeffreys' prior could improve on the coverage of true LV scores. We also propose to combine BC and the use of Jeffreys' prior, in the hope to obtain interval estimates of LV scores that better characterize both the measurement and sampling error.

4.2. *Jeffreys' Prior*

As indicated by the Bernstein–von Mises theorem, the posterior probability corresponding to a quantile-based Bayesian interval estimator usually matches the frequentist coverage probability of θ_0 up to a residual term of order $m^{-1/2}$, in which m is the number of test items. Aiming at better coverage of θ_0 while retaining the framework of Bayesian scoring, we seek for an approximate higher-order matching between the two probabilities. In general, when the parameter is unidimensional and the data are continuous, Jeffreys' prior is typically *first-order matching*: The difference between the posterior and coverage probabilities converges to 0 at the rate m^{-1}, faster than the usual rate $m^{-1/2}$ (Welch & Peers, 1963; Datta & Mukerjee, 2004). For discrete data, randomization may be necessary for Jeffreys' prior to be exactly first-order matching (Rousseau, 2000). However, our simulation results (see Sect. 5) suggest that using Jeffreys' prior for response pattern scoring often leads to good frequentist coverage of the true LV scores even without randomization. In addition, when the items are interchangeable, i.e., with identical item parameters, the estimation of θ_0 is isomorphic to the estimation of a binomial proportion. For the latter problem, interval estimation based on Jeffreys' prior has been advocated in the literature (Brown, Cai, & DasGupta, 2001; 2002).

Consider a random item response vector $\mathbf{Y}_0$ that follows the conditional probability mass function $f(\mathbf{y}_0|\theta_0; \boldsymbol{\beta}) = \prod_{j=1}^{m} f_j(y_{0j}|\theta_0; \boldsymbol{\beta}_j)$ (i.e., Eq. 1). Let

$$\mathcal{I}(\theta; \boldsymbol{\beta}) = E_{\theta,\boldsymbol{\beta}}^{\mathbf{Y}_0}\left\{\left[\frac{d \log f(\mathbf{Y}_0|\theta; \boldsymbol{\beta})}{\mathrm{d}\theta}\right]^2\right\} = -E_{\theta,\boldsymbol{\beta}}^{\mathbf{Y}_0}\left[\frac{d^2 \log f(\mathbf{Y}_0|\theta; \boldsymbol{\beta})}{\mathrm{d}\theta^2}\right] \tag{15}$$

be the Fisher information with respect to θ, which is typically referred to as the *test information*, in which $E_{\theta,\boldsymbol{\beta}}^{\mathbf{Y}_0}$ denotes the expectation taken with respect to $f(\cdot|\theta; \boldsymbol{\beta})$. Jeffreys' prior is proportional to the square root of the test information (Jeffreys, 1946):

$$\phi^J(\theta; \boldsymbol{\beta}) \propto \sqrt{\mathcal{I}(\theta; \boldsymbol{\beta})}. \tag{16}$$

Test information (Eq. 15) for various commonly used IRT models can be found in, for example, Baker and Kim (2004).

4.3. *Combining Bootstrap Calibration and Jeffreys' Prior*

It has already been demonstrated that directly substituting the ML estimates of the item parameters $\hat{\boldsymbol{\beta}}$ for their true values $\boldsymbol{\beta}_0$ in the posterior calculation (Eqs. 3, 5), i.e., the plug-in method, could compromise the credibility of the resulting interval estimates of LV scores when the measurement model is calibrated using a small sample. Because the BC method has been applied successfully in response pattern scoring with a standard normal prior, we anticipate that it could also enhance the quality of Jeffreys' prior interval estimates. It is emphasized that we

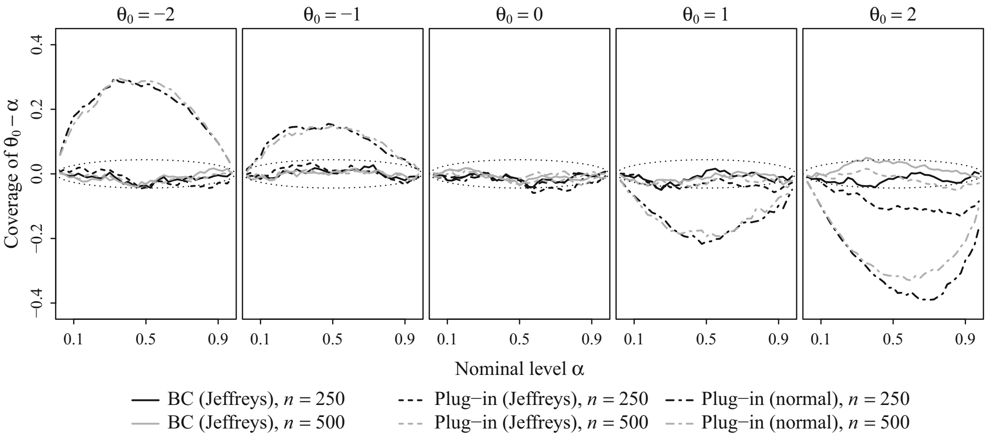

FIGURE 5.

Empirical coverage of θ_0 for one-sided interval estimates of LV scores (sample size $n = 250$ and 500, test length $m = 36$, normal and Jeffreys' priors). Simulation results for different true LV score levels ($\theta_0 = -2, -1, 0, 1$, and 2) are displayed in separate panels. In *each panel*, the x-axis is the nominal α level, and the y-axis is the discrepancy between the empirical coverage and the nominal α level. Results under the two sample size conditions are depicted in different colors: *black* for $n = 250$ and *gray* for $n = 500$. The empirical coverage values for different methods are shown in *different line* types. The *dotted curves* delineate a 95% pointwise confidence envelope for the coverage probability based on normal approximation.

only use Jeffreys' prior in the scoring stage, not the calibration stage. This hybrid use of priors implicitly poses an assumption that the calibration sample was drawn from a population in which the LV follows a particular distribution (such as a standard normal distribution), whereas the response data to be scored need not come from the same population.

5. Simulation 2

The finite-sample performance of Jeffreys' prior intervals is examined by additional Monte Carlo simulations. The simulation design and software configuration were the same as is described in Sect. 3, with the exception that we used Jeffreys' prior instead of $\mathcal{N}(0, 1)$ in the scoring stage. Pilot simulation studies suggested that the posterior distributions derived from Jeffreys' prior are often substantially wider than those derived from a standard normal prior, especially when the true θ_0 value that underlies the scoring data is extreme. In order to maintain the accuracy of numerical integration, we specified 101 Gauss–Legendre quadrature points rescaled to the interval $[-10, 10]$ when approximating the posterior cdf (Eq. 5) and quantile function (Eq. 6).

We first contrast the two types of Jeffreys' prior intervals (plug-in and BC) with the normal prior plug-in interval by means of the empirical coverage of the true LV score θ_0; the simulation results are summarized in Fig. 5. As we expect, the coverage is substantially improved by using Jeffreys' prior, especially when θ_0 is away from 0. When the calibration sample size is small ($n = 250$), however, the coverage of the plug-in Jeffreys' prior intervals can be significantly lower than the nominal level when $\theta_0 \geq 1$; in Fig. 5, it is visualized by the lines' falling below the 95% normal approximation confidence envelope of coverage. Compared to the plug-in method, BC results in a closer match to the nominal coverage level throughout most θ_0 and α levels: The improvement is quite notable (about 0.1) sometimes. As the calibration sample size increases to 500, the two methods based on Jeffreys' prior yield virtually the same results.

The predictive coverage of plug-in and BC-based Jeffreys' prior interval estimators are compared in Fig. 6. Note that Jeffreys' prior was also used in calculating the empirical predictive

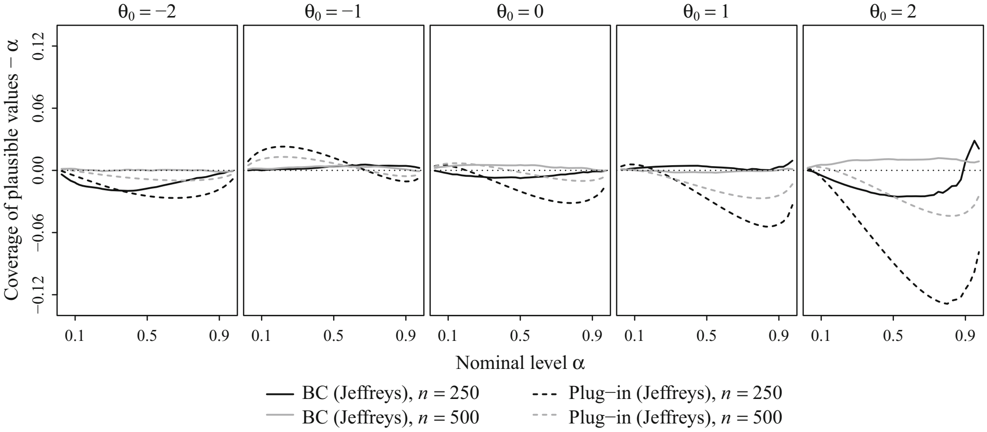

FIGURE 6.
Empirical predictive coverage for one-sided interval estimates of LV scores (sample size $n = 250$ and 500, test length $m = 36$, Jeffreys' prior). Simulation results for different true LV score levels ($\theta_0 = -2, -1, 0, 1$, and 2) are displayed in separate panels. In *each panel*, the x-axis is the nominal α level, and the y-axis is the discrepancy between the empirical predictive coverage and the nominal α level. Results under the two sample size conditions are depicted in different colors: *black* for $n = 250$ and *gray* for $n = 500$. *Solid* and *dashed lines* represent the empirical coverage values for bootstrap calibration (BC) and plug-in methods, respectively.

coverage. In the small calibration sample condition ($n = 250$), the BC-based interval attains a close match to the nominal predictive coverage at most θ_0 and α levels, similar to what we found in the first simulation study (Sect. 3). For BC, we also observe a steep increase in the coverage of $\theta_0 = 2$ as the nominal level exceeds 0.8; we conjecture that it is traceable to the limited accuracy of numerical integration. The predictive coverage of the plug-in interval can be more than 0.12 off from the nominal level when $\theta_0 = 2$. Even when $n = 500$, it still fall short of the nominal level: At $\theta_0 = 2$, the difference is more than 0.04 for α between 0.7 and 0.9.

In summary, if we are interested in recovering the true individual LV score from an observed response pattern, then interval estimates obtained using Jeffreys' prior is more trustworthy. When the calibration sample is also small, we prefer BC over the plug-in method for the reason that empirically both the predictive coverage and the coverage of θ_0 are in favor of the former. However, if the LV indeed follows a normal distribution in the population and we are only interested in describing as a population characteristic the distribution of the LV conditional on an observed response pattern, then we should still calculate the normal prior intervals together with BC.

6. Empirical Examples

6.1. Example 1: A 12th-Grade Science Test

The first example concerns a 12th-grade science test covering the subjects of biology, chemistry, and physics. The data can be accessed from both TESTFACT (Wood et al., 2003) and the R package `mirt`. The test is composed of 32 items ($m = 32$), and there are a total of 600 observed response patterns in the data set. We dichotomously scored the original responses using the provided answer key. Three hundred response patterns were randomly selected as the calibration sample ($n = 300$), in which we fitted a 3PL model with all pseudo-guessing parameters constrained equal, i.e., the same model as was used in our simulation study.

The remaining 300 response patterns form the scoring sample. For each response pattern, we computed the plug-in and BC-based estimates of the quantiles (i.e., the upper limits of the one-

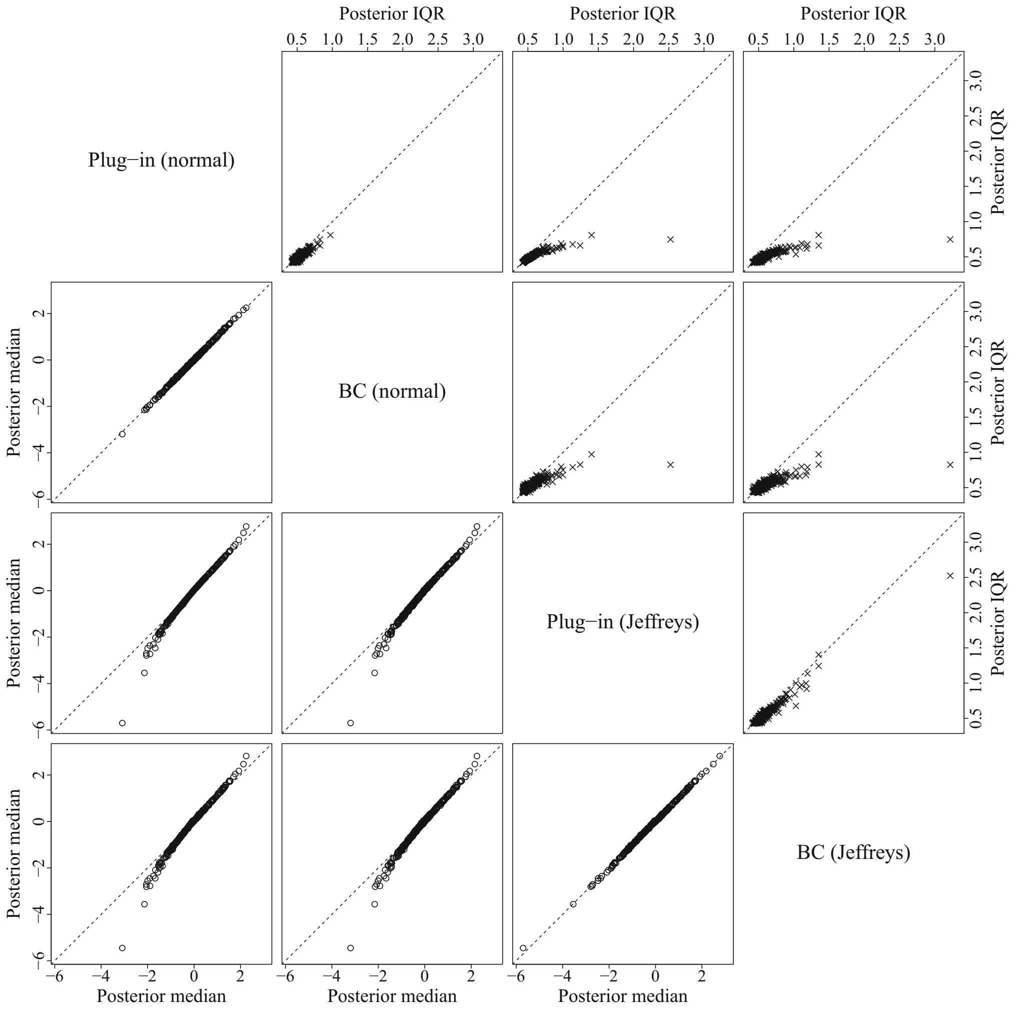

FIGURE 7.
Point estimates (posterior median) and precision estimates (posterior IQR) of LV scores for 300 randomly selected response patterns to the 32-item science assessment test. The four candidate methods are based on a cross-classification of two factors: standard normal versus Jeffreys' priors, and plug-in versus BC estimates. The posterior medians (*circles*) are plotted in the *lower-triangular panels*, while the posterior IQRs (*crosses*) are in the *upper-triangular panels*. The *diagonal dashed line* in *each panel* indicates the identity of the two coordinates.

sided interval estimates) of the posterior distributions derived from both the standard normal and Jeffreys' priors. Three nominal α levels were considered: $\alpha = 0.25$, 0.5, and 0.75: The posterior median (i.e., $\alpha = 0.5$) serves as a point estimator of the LV score (similar to the EAP score which is the posterior mean), and the difference between the 0.75th and 0.25th quantiles yields the posterior inter-quartile range (IQR) which measures the variability of the score. In Fig. 7, each pair of the four candidate methods was contrasted; results for the posterior median and IQR were located in the lower- and upper-triangular panels, respectively.

The standard normal prior leads to less variable posterior medians and smaller IQRs than the Jeffreys' prior does, especially when the estimated scores are extreme. The plug-in and BC-based methods yield essentially identical estimates of the posterior medians, provided the same prior is in use. The BC-based predictive distributions are often wider than the plug-in ones: For 86 and

97 out of 300 response patterns under the normal and Jeffreys' priors, respectively, the BC-based IQR estimates are more than 10% larger than the corresponding plug-in estimates. Sometimes, BC yields slightly narrower predictive distributions, but the reduction in IQR is seldom beyond 5% (3 cases under the normal prior and 12 cases under the Jeffreys' prior).

6.2. *Example 2: An Attitude Survey About Science and Technology*

In the second example, we analyze a four-item test assessing general attitude toward science and technology. The data set was collected as a part of the 1992 Euro-Barometer Survey (Karlheinz and Melich, 1992) and is available in the R packages `mirt` and `ltm` (Rizopoulos, 2006). The four test items are measured on the same four-point likert scale ranging from "strongly disagree" to "strongly agree." We fitted a graded response model using the 392 complete responses contained in the data set ($n = 392$). Plug-in and BC-based posterior medians and IQRs were calculated for all $4^4 = 256$ possible response patterns to the four ordinal items. The results are presented in Fig. 8, a scatterplot matrix arranged similar to Fig. 7.

A pattern similar to the first example is obtained. The shrinkage imposed by the standard normal prior is salient because the test is extremely short ($m = 4$). Under the same prior, the BC-based and plug-in estimates of the posterior medians are virtually the same. Most often, the BC-based IQR estimates are larger than the plug-in estimates: When the normal prior is used, more than 10% increase in IQR is observed for 48 out of 256 response patterns after applying BC; the number of affected response patterns becomes 101 when Jeffreys' prior is used. For 11 response patterns under the normal prior and 6 under Jeffreys' prior, the BC-based posterior IQRs are more than 5% smaller than the plug-in ones. We conclude from the two examples that the plug-in posterior may substantially overstate measurement precision when the calibration sample size is small.

7. Discussion

It is a common practice to use estimated IRT model parameters in scoring individual response patterns; however, it has been warned that the plug-in method brings extra uncertainty into the LV score estimates, for the reason that the estimated model parameters are subject to sampling error. In this article, we introduced a Monte Carlo method, namely BC, for handling the propagation of sampling error from model calibration to scoring. Asymptotically, the BC-based one-sided interval estimator of LV scores usually achieves higher-order accuracy in predictive coverage, a criterion that quantifies the impact of sampling error in the Bayesian scoring framework, as compared to the widely used plug-in interval estimator. Monte Carlo simulations were conducted to confirm the advantage of BC in small sample calibrations. We also recommend the use of Jeffreys' prior for a better recovery of the true LV scores when the test is short and/or the true LV score is extreme. By combining the use of BC and Jeffreys' prior, we were able to secure valid statistical inference for the true LV scores even when the model parameters were estimated from a small calibration sample.

There are several limitations and extensions to be addressed by future research.

First, the current application of BC is limited to unidimensional IRT models; generalizations to multidimensional IRT models should be considered in follow-up studies. If we are interested in marginal interval estimates for one latent dimension at a time, the methods discussed in the current work still apply, except that the integrations appearing in the posterior cdf (Eq. 5) are now multidimensional. Quadrature-based numerical integration becomes inefficient as the dimensionality grows; as a workaround, the double bootstrap procedure (Algorithm 2 in Beran, 1990; see a brief discussion in Sect. 2.4) can be counted on. Simultaneous set estimation (which generalizes interval estimation) for more than one coordinate of the LV is slightly more complicated.

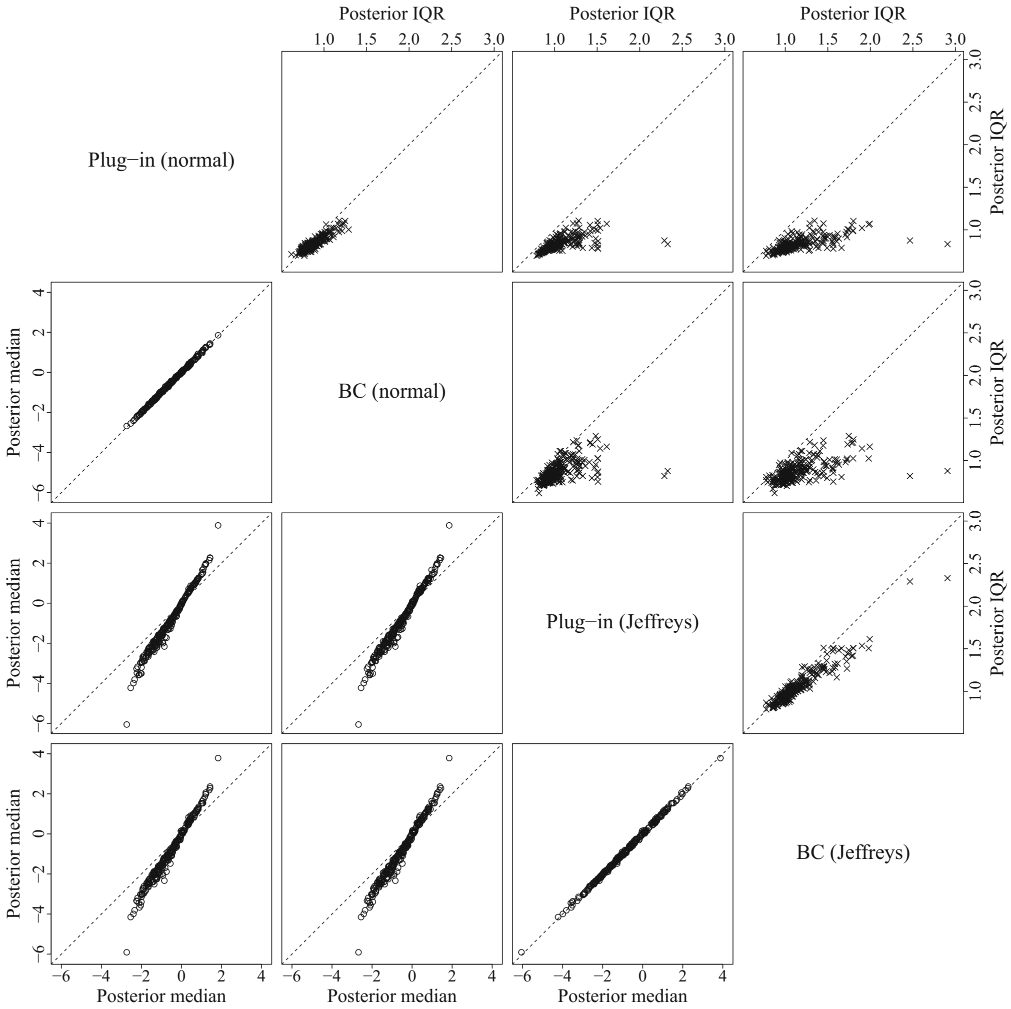

FIGURE 8.
Point estimates (posterior median) and precision estimates (posterior IQR) of LV scores for all $4^4 = 256$ possible response patterns to the four-item attitude survey about science and technology. The four candidate methods are based on a cross-classification of two factors: standard normal versus Jeffreys' priors, and plug-in versus BC estimates. The posterior medians (*circles*) are plotted in the *lower-triangular panels*, while the posterior IQRs (*crosses*) are in the *upper-triangular panels*. The *diagonal dashed line* in *each panel* indicates the identity of the two coordinates.

In a multidimensional space, the shape of a set estimator is often specified via the so-called root variable (Beran, 1990), whose quantile determines the boundary of the set estimate. For instance, if we take the minus posterior density function as the root variable, the resulting set estimator is the highest posterior density region. BC-based set estimates can be obtained by applying a Monte Carlo procedure similar to Algorithm 1 to the root variable, instead of the LV itself. In addition, BC can be applied to other types of generalized LV measurement models (Bartholomew & Knott, 1999; Muthén, 2002; Skrondal and Rabe-Hesketh, 2004) such as linear-normal or Poisson factor analysis.

Second, we mainly focus on the derivation and implementation of BC in the current work; a comprehensive simulation comparison among existing methods is left for future research. Methodological studies belonging to the latter category help applied researchers and practitioners to select the most appropriate method in their own analyses. Apart from the predictive coverage and the coverage of the true LV score that was used in this article, practical matters such as execution time and various tuning aspects of the computation are also subject to evaluation. The theoretical framework of Bayesian scoring and predictive inference facilitates large-scale simulation comparisons among most Bayesian, fiducial, and frequentist inferential methods for IRT scale scores.

Third, the computational efficiency of BC can be improved. Taking advantage of the multicore architecture of modern computers, computations in different bootstrap samples (Lines 4–6 of Algorithm 1) can be parallelized in separate processing units. In addition, despite simplicity in implementation, numerical integration with the same set of quadrature points fails to account for the fact that posterior distributions corresponding to different response patterns could differ substantially in both location and variability. Adaptive Gaussian quadrature similar to Schilling and Bock (2005) and Haberman (2006) can be invoked to obtain more efficient and accurate approximations to the marginal probability, posterior cdf, and posterior quantiles.

Finally, in many educational and psychological research, it is the associations among latent constructs, rather than individual scores, that are of major interest. Statistical inference of structural models (e.g., multiple regression, path analysis, multilevel modeling, etc.) using estimated LV scores is beyond the scope of this article; however, we remark that the naive use of estimated LV scores as the input data for structural models is an error-prone practice, because the estimated scores are contaminated by measurement and sampling error. We refer to Carroll et al. (2006) for a comprehensive overview of handling measurement error in general linear and nonlinear statistical models.

Acknowledgments

We are grateful to Dr. David Thissen from the Department of Psychology at the University of North Carolina at Chapel Hill for his advice and feedback on this paper.

Appendix A. Theoretical Details

A.1. The ML Estimator

Let $\mathbf{f}(\boldsymbol{\xi})$ be the vector of all response pattern probabilities (with elements $f(\mathbf{y}_i;\boldsymbol{\xi})$ defined in Eq. 4), and $\mathbf{p}$ denote the corresponding observed proportions. Consider a calibration sample of n i.i.d. item response patterns $\mathbf{Y}_1,\dots,\mathbf{Y}_n$, each of which is generated from the unidimensional IRT model characterized by $\mathbf{f}(\boldsymbol{\xi}_0)$. In some neighborhood of $\boldsymbol{\xi}_0$, denoted U_0, suppose that standard regularity conditions (Birch, 1964; Bishop et al., 1975; Cox, 1984) hold, and that $\mathbf{f}(\boldsymbol{\xi})$ has continuous fourth partial derivatives, then there exists a four-time continuously differentiable function $\boldsymbol{\xi}(\cdot)$ such that $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(\mathbf{f}(\boldsymbol{\xi}_0))$ and the ML estimator $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}(\mathbf{p})$ for all admissible probability vector $\mathbf{p} \in N_0$ where N_0 is some neighborhood of $\mathbf{f}(\boldsymbol{\xi}_0)$. As n tends to infinity, $\mathbf{p}$ concentrates on N_0 exponentially fast: It follows from Hoeffding's (1963) inequality that

$$P_{\boldsymbol{\xi}_0}^{\mathbf{Y}}\{\mathbf{p} \notin N_0\} \le B\exp(-cn) \tag{17}$$

for some positive constants B and c.

The following equation, which resembles Beran's (1990) Assumption A(4), can be verified by a Taylor series expansion argument similar to Lehmann and Casella (1998, p. 430):

$$E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}\left[\left(\boldsymbol{\xi}(\mathbf{p})-\boldsymbol{\xi}_0\right)^{\mathbf{r}}\mathbb{I}\{\mathbf{p}\in N_0\}\right]=\sum_{s=\lfloor(|\mathbf{r}|+1)/2\rfloor}^{2} n^{-s}b_{\mathbf{r},s}(\boldsymbol{\xi}_0)+o(n^{-2}),\quad |\mathbf{r}|=1,\ldots,4. \tag{18}$$

In Eq. 18, the Q-tuple of nonnegative integers $\mathbf{r}=(r_1,\ldots,r_q)^\top$ serves as a multi-index such that $\boldsymbol{\xi}^{\mathbf{r}}=\xi_1^{r_1}\cdots\xi_q^{r_q}$ for any q-dimensional vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_q)^\top$, and $|\mathbf{r}|=r_1+\cdots+r_q$. $\mathbb{I}\{\cdot\}$ denotes the indicator function. The functions $b_{\mathbf{r},s}(\boldsymbol{\xi}_0)$ for all $\mathbf{r}$ and s are related to the partial derivatives of $\boldsymbol{\xi}(\cdot)$ and the moments of $\mathbf{p}$ up to the fourth order. Equation 18 indicates that the first and second moments of the truncated ML estimator are of order $O(n^{-1})$, and that the third and fourth moments are of order $O(n^{-2})$.

A.2. The Plug-in Method

Let the cdf $G(\cdot;\boldsymbol{\xi})$ be continuous for all $\boldsymbol{\xi}$ in the closure of the parameter space. In particular, we assume that $G(\theta;\boldsymbol{\xi})$ is strictly monotonic in θ and four-time continuously differentiable with respect to both θ and $\boldsymbol{\xi}$ for all $\theta\in\mathbb{R}$ and $\boldsymbol{\xi}\in U_0$. These assumptions imply that $C(\alpha;\boldsymbol{\xi},\boldsymbol{\xi}_0)=G(G^{-1}(\alpha;\boldsymbol{\xi});\boldsymbol{\xi}_0)$ is always defined and has continuous fourth partial derivatives with respect to $\boldsymbol{\xi}$ for any $\boldsymbol{\xi}\in U_0$. Then, $C(\alpha;\boldsymbol{\xi},\boldsymbol{\xi}_0)$ has the following Taylor series expansion at $\boldsymbol{\xi}=\boldsymbol{\xi}_0$:

$$C(\alpha;\boldsymbol{\xi},\boldsymbol{\xi}_0)=\alpha+\sum_{|\mathbf{r}|=1}^{3}\frac{\partial^{|\mathbf{r}|}C}{\partial\boldsymbol{\xi}^{\mathbf{r}}}(\alpha,\boldsymbol{\xi}_0,\boldsymbol{\xi}_0)\frac{(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^{\mathbf{r}}}{\mathbf{r}!}+\sum_{|\mathbf{r}|=4}\frac{\partial^4C}{\partial\boldsymbol{\xi}^{\mathbf{r}}}(\alpha,\bar{\boldsymbol{\xi}},\boldsymbol{\xi}_0)\frac{(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^{\mathbf{r}}}{\mathbf{r}!} \tag{19}$$

in which $\mathbf{r}!=r_1!\cdots r_Q!$, and $\bar{\boldsymbol{\xi}}$ lies between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_0$. Note that $C(\alpha;\boldsymbol{\xi}_0)=E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}[C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0)]$ is essentially $E^{\mathbf{Y}}_{\boldsymbol{\xi}_0}\left[C(\alpha;\hat{\boldsymbol{\xi}},\boldsymbol{\xi}_0)\mathbb{I}\{\mathbf{p}\in N_0\}\right]$ plus an exponentially small term because of Eq. 17 and the boundedness of $C(\alpha;\boldsymbol{\xi},\boldsymbol{\xi}_0)$. It follows from Eqs. 18 and 19 that

$$C(\alpha;\boldsymbol{\xi}_0)=\alpha+n^{-1}d(\alpha,\boldsymbol{\xi}_0)+O(n^{-2}), \tag{20}$$

in which

$$d(\alpha,\boldsymbol{\xi}_0)=\sum_{|\mathbf{r}|=1}^{2}\frac{\partial^{|\mathbf{r}|}C}{\partial\boldsymbol{\xi}^{\mathbf{r}}}(\alpha,\boldsymbol{\xi}_0,\boldsymbol{\xi}_0)\frac{b_{\mathbf{r},1}(\boldsymbol{\xi}_0)}{\mathbf{r}!}. \tag{21}$$

A.3. Predictive Calibration

Analogously, expanding $C(C^{-1}(\alpha;\boldsymbol{\xi}_0);\boldsymbol{\xi},\boldsymbol{\xi}_0)$ at $\boldsymbol{\xi}=\boldsymbol{\xi}_0$ and taking expectation yield

$$C^{-1}(\alpha;\boldsymbol{\xi}_0)=\alpha-n^{-1}d(\alpha,\boldsymbol{\xi}_0)+O(n^{-2}). \tag{22}$$

The assumptions we made imply that Eq. 22 still holds with the same $d(\cdot,\cdot)$ if we replace $\boldsymbol{\xi}_0$ by $\boldsymbol{\xi}$ in some neighborhood $V_0\subset U_0$,[8] and $\sup_{\boldsymbol{\xi}\in V_0}|C^{-1}(\alpha;\boldsymbol{\xi})-\alpha+n^{-1}d(\alpha,\boldsymbol{\xi})|=O(n^{-2})$. Plugging in the expansion of $d(\alpha,\boldsymbol{\xi})$ at $\boldsymbol{\xi}=\boldsymbol{\xi}_0$, we obtain

$$C^{-1}(\alpha;\boldsymbol{\xi})=\alpha-n^{-1}d(\alpha,\boldsymbol{\xi})+O(n^{-2})$$

[8] N_0 and the truncated ML estimator remain unchanged.

$$
\begin{aligned}
&= \alpha - n^{-1}\left\{ d(\alpha, \boldsymbol{\xi}_0) + (\boldsymbol{\xi} - \boldsymbol{\xi}_0)^{\top} \frac{\partial d}{\partial \boldsymbol{\xi}}(\alpha, \boldsymbol{\xi}_0) + \sum_{|\mathbf{r}|=2} \frac{\partial^2 d}{\partial \boldsymbol{\xi}^{\mathbf{r}}}(\alpha, \breve{\boldsymbol{\xi}}) \frac{(\boldsymbol{\xi} - \boldsymbol{\xi}_0)^{\mathbf{r}}}{\mathbf{r}!} \right\} \\
&+ O(n^{-2}), \qquad (23)
\end{aligned}
$$

in which $\breve{\boldsymbol{\xi}}$ falls between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_0$.

We now expand $C(C^{-1}(\alpha; \boldsymbol{\xi}); \boldsymbol{\xi}, \boldsymbol{\xi}_0)$ with respect to the first argument at $C^{-1}(\alpha; \boldsymbol{\xi}_0)$ and then further expand $\frac{\partial C}{\partial \alpha}(C^{-1}(\alpha; \boldsymbol{\xi}_0), \boldsymbol{\xi}, \boldsymbol{\xi}_0)$ with respect to the second argument at $\boldsymbol{\xi}_0$:

$$
\begin{aligned}
&C(C^{-1}(\alpha; \boldsymbol{\xi}); \boldsymbol{\xi}, \boldsymbol{\xi}_0) \\
&= C(C^{-1}(\alpha; \boldsymbol{\xi}_0); \boldsymbol{\xi}, \boldsymbol{\xi}_0) + \frac{\partial C}{\partial \alpha}(C^{-1}(\alpha; \boldsymbol{\xi}_0), \boldsymbol{\xi}, \boldsymbol{\xi}_0)[C^{-1}(\alpha; \boldsymbol{\xi}) - C^{-1}(\alpha; \boldsymbol{\xi}_0)] \\
&+ \frac{1}{2}\frac{\partial^2 C}{\partial \alpha^2}(\bar{\alpha}, \boldsymbol{\xi}, \boldsymbol{\xi}_0)[C^{-1}(\alpha; \boldsymbol{\xi}) - C^{-1}(\alpha; \boldsymbol{\xi}_0)]^2 \\
&= C(C^{-1}(\alpha; \boldsymbol{\xi}_0); \boldsymbol{\xi}, \boldsymbol{\xi}_0) + \left\{ \frac{\partial C}{\partial \alpha}(C^{-1}(\alpha; \boldsymbol{\xi}_0), \boldsymbol{\xi}_0, \boldsymbol{\xi}_0) + (\boldsymbol{\xi} - \boldsymbol{\xi}_0)^{\top} \left[\frac{\partial^2 C}{\partial \alpha \partial \boldsymbol{\xi}}(C^{-1}(\alpha; \boldsymbol{\xi}_0), \tilde{\boldsymbol{\xi}}, \boldsymbol{\xi}_0) \right] \right\} \\
&\cdot [C^{-1}(\alpha; \boldsymbol{\xi}) - C^{-1}(\alpha; \boldsymbol{\xi}_0)] + \frac{1}{2}\frac{\partial^2 C}{\partial \alpha^2}(\bar{\alpha}, \boldsymbol{\xi}, \boldsymbol{\xi}_0)[C^{-1}(\alpha; \boldsymbol{\xi}) - C^{-1}(\alpha; \boldsymbol{\xi}_0)]^2, \qquad (24)
\end{aligned}
$$

in which $\bar{\alpha}$ falls between $C^{-1}(\alpha; \boldsymbol{\xi}_0)$ and $C^{-1}(\alpha; \boldsymbol{\xi})$, and $\tilde{\boldsymbol{\xi}}$ fall between $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}$. By combining Eqs. 23 with 24 and taking expectation, we conclude that the calibrated predictive coverage matches α up to an $O(n^{-2})$ error term.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

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

Barndorff-Nielsen, O. E., & Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli*, *2*(4), 319–340.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Edward Arnold (Kendall's Library of Statistics 7).

Beran, R. (1990). Calibrating prediction regions. *Journal of the American Statistical Association*, *85*(411), 715–723.

Birch, M. W. (1964). A new proof of the pearson-fisher theorem. *The Annals of Mathematical Statistics*, *35*(2), 817–824.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bishop, Y., Fienberg, S., & Holland, P. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: The MIT Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*(2), 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Bolt, D. M. (2005). Limited and full-information IRT estimation. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). New Jersey: Lawrence-Erlbaum.

Brent, R. P. (1973). Some efficient algorithms for solving systems of nonlinear equations. *SIAM Journal on Numerical Analysis*, *10*(2), 327–344.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–117.

Brown, L. D., Cai, T. T., & Dasgupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics*, *30*(1), 160–201.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for windows*. Lincolnwood, IL: Scientific Software International.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca-Raton, FL: CRC press.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software,48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*(1), 37–52.

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280–291.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Cox, C. (1984). An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *The American Statistician*, *38*(4), 283–287.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), 248–275.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100.

Datta, G. S., & Mukerjee, R. (2004). *Probability matching priors: Higher order asymptotics*. New York: Springer.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Fonseca, G., Giummolè, F., & Vidoni, P. (2014). Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, *84*(2), 373–383.

Haberman, S. J. (2006). *Adaptive quadrature for item response models*. Technical report no. 06-29, Educational Testing Service, Princeton, NJ.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*. http://pareonline.net/getvn.asp?v=17&n=1

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software manual]*. Chapel Hill, NC: Vector Psychometric Group.

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J.-S., et al. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*(4), 595–607.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. In *Proceedings of the royal society of London A: Mathematical, physical and engineering sciences* (Vol. 186, pp. 453–461).

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York: Wiley.

Le Cam, L., & Yang, G. L. (2000). *Asymptotics in statistics: Some basic concepts* (2nd ed.). New York: Springer.

Lehmann, E., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Berlin: Springer.

Liu, Y., & Hannig, J. (2016). Generalized fiducial inference for binary logistic item response models. *Psychometrika*, *81*(2), 290–324.

Liu, Y., & Hannig, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrika*. doi:10.1007/s11336-017-9554-0.

Magnus, B. E., Liu, Y., He, J., Quinn, H., Thissen, D., Gross, H. E., et al. (2016). Mode effects between computer self-administration and telephone interviewer-administration of the PROMIS pediatric measures, self-and proxy report. *Quality of Life Research*, *25*, 1655–1665.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Mislevy, R. J., Wingersky, M., & Sheehan, K. M. (1993). *Dealing with uncertainty about item parameters: Expected response functions*. Technical report no. 94-28, Educational Testing Service, Princeton, NJ.

Muenks, K., Wigfield, A., Yang, J. S., & O'Neal, C. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement. *Journal of Educational Psychology, 109*(5), 599–620.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47–73.

Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*, 24–40.

Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2014). Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement*, *74*(4), 697–712.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178.

R Core Team. (2016). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria. Retrieved from https://www.R-project.org/.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Annals of the Institute of Statistical Mathematics*, *52*(1), 28–42.

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), 3–38.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. In *Psychometrika monograph no. 17*. Richmond, VA: Psychometric Society.

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: Statistical tools* (pp. 127–150). Boca Raton: CRC Press.

San Martín, E., & De Boeck, P. (2015). What do you mean by a difficult item? On the interpretation of the difficulty parameter in a Rasch model. In R. Millsap, D. Bolt, L. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (pp. 1–14). Berlin: Springer.

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika*, *78*(2), 341–379.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533–555.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Skrondal, A., & Rabe-Hasketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall. (Interdisciplinary Statistics Series).

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). London: Sage Publications.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, *85*(4), 949–953.

Vidoni, P. (2009). Improved prediction intervals and distribution functions. *Scandinavian Journal of Statistics*, *36*(4), 735–748.

Welch, B., & Peers, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *25*(2), 318–329.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281–301.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2003). *TESTFACT 4 for windows: Test scoring, item statistics, and full-information item factor analysis [Computer software]*. Lincolnwood, IL: Scientific Software International.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290.

Manuscript Received: 21 SEP 2016
Final Version Received: 22 MAY 2017
Published Online Date: 6 SEP 2017