Article

Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 6, pp. 547–578
DOI: 10.3102/1076998615606112

IRT Item Parameter Recovery With Marginal Maximum Likelihood Estimation Using Loglinear Smoothing Models

Jodi M. Casabianca
The University of Texas at Austin

Charles Lewis
Fordham University

Loglinear smoothing (LLS) estimates the latent trait distribution while making fewer assumptions about its form and maintaining parsimony, thus leading to more precise item response theory (IRT) item parameter estimates than standard marginal maximum likelihood (MML). This article provides the expectation-maximization algorithm for MML estimation with LLS embedded and compares LLS to other latent trait distribution specifications, a fixed normal distribution, and the empirical histogram solution, in terms of IRT item parameter recovery. Simulation study results using a 3-parameter logistic model reveal that LLS models matching four or five moments are optimal in most cases. Examples with empirical data compare LLS to these approaches as well as Ramsay-curve IRT.

Keywords: *item response theory; nonnormal distributions; marginal maximum likelihood; latent trait distribution; quadrature; loglinear smoothing; moments; EM algorithm*

Unidimensional item response theory (IRT; Lord & Novick, 1968) models play an important role in present-day educational measurement and testing. With marginal maximum likelihood (MML) estimation of IRT item parameters (Bock & Aitkin, 1981; Bock & Lieberman, 1970), the IRT model is expanded to include a latent trait distribution, denoted here by $g(\theta)$. The latent trait is considered to be a nuisance parameter in MML and can be eliminated from the likelihood by integrating or marginalizing over it, a strategy that goes back to Neyman and Scott (1948). Item parameter recovery studies have shown that there is item parameter estimation error when the true $g(\theta)$ is nonnormal and a normal distribution is specified (Boulet, 1996; Stone, 1992; Swaminathan & Gifford, 1983; Woods & Lin, 2009; Woods & Thissen, 2006; Yamamoto & Muraki, 1991). Many applications of IRT use item parameter estimates from a calibration sample in subsequent analyses, including trait estimation or equating. Given that there is a relationship between the amount of systematic and random error in item parameter estimates and the amount of measurement error in estimates from these

subsequent analyses (Casabianca, 2011; Casabianca & Lewis, 2011a; Tsutakawa & Johnson, 1990; Zhang, 2011; Zhang, Xie, Song, & Lu, 2011), an interest of psychometricians is to obtain the most accurate and precise item parameter estimates possible by calibrating with large samples and the appropriate estimation specifications.

Flexibility in model specification for $g(\theta)$ is important. Specifically, the model for $g(\theta)$ should be able to accommodate nonnormality while also maintaining a level of parsimony. Current practice for specifying the latent trait distribution assumes a normal distribution by default, and this is true in Parscale 4 (Muraki & Bock, 2003), Bilog-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003), and more current software such as flexMIRT Version 2.0 (Cai, 2013). Although it is not always apparent, there are many situations in educational and psychological testing, where the population of test takers may have an approximately normal observed test score distribution but a nonnormal distribution for the latent trait (Lord, 1953; Lord & Novick, 1968). An example in educational testing where the latent trait may be inherently nonnormal is a group of examinees with disabilities taking a K–12 alternate assessment. Measurement of constructs in the field of clinical and personality psychology often involve nonnormal traits—for example, psychoticism, tends to be positively skewed (Eysenck & Eysenck, 1991; Matthews, Deary, & Whiteman, 2003, p. 22).

Generally, parsimony is a major consideration in statistical modeling. It is especially important in psychometrics with the usage of relatively more complex models such as the unidimensional 3-parameter logistic (3PL) IRT model, multidimensional IRT models, and cognitive diagnostic models. These models involve many estimable parameters and are typically associated with identifiability issues (Haberman, 2005). Therefore, flexible yet parsimonious options for estimation should be considered in these cases, as the ideal model for $g(\theta)$ sits at the optimal point of the bias–variance trade-off.

The most common *fixed* specification for $g(\theta)$ is a standard normal distribution. A discrete approximation to the normal distribution is used for the purposes of numerical integration (Bock & Lieberman, 1970; Mislevy, 1984). The most common *parametric* approach is to estimate the parameters of the normal distribution (Andersen & Madsen, 1977; Mislevy, 1984; Rigdon & Tsutakawa, 1983; Sanathanan & Blumenthal, 1978). The most common *nonparametric* approach to model $g(\theta)$, often called the Bock–Aitkin or the empirical histogram (EH) solution (Bock & Aitkin, 1981; Mislevy & Bock, 1985; Rigdon & Tsutakawa, 1983), estimates $Q - 1$ distributional parameters that characterize a discrete $g(\theta)$ using latent class probability estimation.

The focus of this article is a *semiparametric* model for $g(\theta)$ via loglinear smoothing (LLS; Holland & Thayer, 1987, 2000) in MML. LLS matches M moments of the original distribution to create a smoothed distribution. Suppose that an LLS model for $g(\theta)$ is embedded within an expectation-maximization

algorithm (EM; Dempster, Laird, & Rubin, 1977; Tsutakawa, 1985) for the MML estimation of item parameters. Given a discretized $g(\theta)$ on Q support or quadrature points, an unsaturated LLS model (with fewer fitted moments than the $Q-1$ points of the distribution or $M < Q-1$) smooths the discrete $g(\theta)$. Conversely, a saturated loglinear model (fitting $Q-1$ moments or $M = Q-1$) will exactly reproduce the original discrete $g(\theta)$.

Other smoothing methods, including LLS, have been used for discrete latent trait and mixture IRT models by Rost and von Davier (1992, 1995), von Davier (2005), and Xu and von Davier (2008). Although these smoothing approaches have been around for quite some time, they were implemented in the context of complex multidimensional discrete latent trait modeling (see, e.g., mdltm software; von Davier, 2005). Furthermore, the algorithms used for estimation of $g(\theta)$ with LLS via the EM algorithm are currently undocumented in the literature and these methods have not been evaluated for use in the MML estimation of item parameters. We believe that for these reasons, LLS algorithms for $g(\theta)$ are also unavailable in popular IRT software packages and are not being utilized. Consequently, our goals for this article are to (i) provide the step-by-step algorithms for LLS in the EM framework, (ii) evaluate LLS as a method for estimating $g(\theta)$ in MMLE of unidimensional IRT models with different values for M and Q, and (iii) compare LLS to other popular approaches.

In the next sections, we review MML estimation of item parameters and provide an overview of the various methods for characterizing $g(\theta)$. We then introduce LLS as a general method for smoothing distributions and detail the specific algorithmic steps to implement the EM algorithm with LLS as the method for estimating $g(\theta)$ in IRT. We report results from an item parameter recovery simulation study performed in the context of the 3PL unidimensional dichotomous IRT model, focusing on item parameter recovery as a function of M and Q. We also present two data examples: one using data from the Programme for International Student Assessment (PISA) mathematics assessment and another from the Maudsley Obsessional Compulsive Inventory (MOCI). Finally, we discuss implications for item parameter estimation and suggestions for implementing LLS in practice.

Characterization of the Latent Trait Distribution in MML Estimation

MML estimation provides estimates of item parameters by assuming that the item response data are obtained from a random sample from a population of latent traits with a certain distribution. With MML, the likelihood is expanded to include $g(\theta)$ and then integrated with respect to θ. The marginal probability of obtaining a specific response pattern $\mathbf{x}_s$ via integration over $g(\theta)$ is

$$P(\mathbf{x}_s) = \int_{-\infty}^{+\infty} P(\mathbf{x}_s|\theta)g(\theta)d\theta = \int_{-\infty}^{+\infty} \left[\prod_{i=1}^{k} P(x_{is}|\theta;\boldsymbol{\varphi})\right] g(\theta)d\theta, \quad (1)$$

where $P(x_{is}|\theta;\boldsymbol{\varphi})$ is an IRT model specifying the conditional probability of obtaining a score x_{is} for item i $(i = 1, \ldots, k)$ from response pattern s $(s = 1, \ldots, S)$, given θ and $\boldsymbol{\varphi}$, a vector of item parameters. The log marginal likelihood for a collection of S response patterns is given by:

$$\ln L = \sum_{s=1}^{S} n_s \ln[P(\mathbf{x}_s)] = \sum_{s=1}^{S} n_s \ln\left\{\int_{-\infty}^{+\infty}\left[\prod_{i=1}^{k} P(x_{is}|\theta;\boldsymbol{\varphi})\right] g(\theta)d\theta\right\}, \tag{2}$$

where n_s is the number of test takers with response pattern s and $\sum_{s=1}^{S} n_s = N$. Note that since each test taker has only one response pattern, the frequencies for each response pattern have a multinomial distribution with parameters, N and $\boldsymbol{P} = [P_1, \ldots, P_S]$, where $P_s = P(\mathbf{x}_s)$ as defined by Equation 1. The MML item parameter estimates are obtained by maximizing $\ln L$ in Equation 2. Numerical integration approximates the integral in Equation 1, as it is not solvable in closed form.

Bock and Lieberman (1970) introduced MML estimation using Gauss–Hermite quadrature to approximate integrals involving the normal distribution (see Stroud & Secrest, 1966). Quadrature utilizes values of a function at a finite set of discrete points to approximate the full area under the curve. A quadrature weight, $W(T_q)$, refers to the height of the function (or in some cases the area of a rectangle) located at quadrature point, T_q. There are several different forms of numerical quadrature, but the standard quadrature in the latent trait context uses fixed points with equal spacing and weights based on a rectangle approximation (i.e., the weight for each point represents the area of a rectangle located at the point).

To approximate the integral in Equation 1, quadrature is used as given by,

$$P(\mathbf{x}_s) \cong \sum_{q=1}^{Q} P(\mathbf{x}_s|T_q)W(T_q) = \sum_{q=1}^{Q}\left[\prod_{i=1}^{k} P(x_{is}|T_q;\boldsymbol{\varphi})\right] W(T_q), \tag{3}$$

where T_q is the quadrature point location and $W(T_q)$ is the weight at quadrature point T_q $(q = 1, \ldots, Q)$. In this equation, the product of the probability of observing item response pattern $\mathbf{x}_s$ at quadrature point T_q and the weight at quadrature point T_q are summed across all Q quadrature points. Inserting the log of Equation 3 into Equation 2 gives the log-likelihood,

$$\ln L = \sum_{s=1}^{S} n_s \ln[P(\mathbf{x}_s)] \cong \sum_{s=1}^{S} n_s \ln\left\{\sum_{q=1}^{Q}\left[\prod_{i=1}^{k} P(x_{is}|T_q;\boldsymbol{\varphi})\right] W(T_q)\right\}. \tag{4}$$

The EH Method for g(θ)

Since θ is unobserved and therefore considered missing, Bock and Aitkin (1981) used the EM algorithm (Dempster, Laird, & Rubin, 1977) to implement

MML estimation of item parameters. In the E-step, expectations of sufficient statistics N_q (the expected number of examinees at T_q) and R_{iq} (the expected number of examinees at T_q responding correctly to item i) are computed so that item parameters may be estimated in the M-step. In the M-step, estimates of the item parameters are computed that maximize the expected log likelihood from the E-step. These estimates are then used as updated quantities to compute the expectations in the next E-step, and so on, until convergence.

The EH method for $g(\theta)$ introduced by Bock and Aitkin (1981) estimates weights $W(T_q)$ iteratively, alongside item parameters, in the EM steps to generate a discrete $g(\theta)$. In the M-step, $W(T_q)$ are estimated with

$$W^{(j+1)}(T_q) = \frac{1}{N}\sum_{s=1}^{S} n_s P^{(j)}(T_q|\mathbf{x}_s) = \frac{1}{N}\sum_{s=1}^{S} n_s \left[\frac{P^{(j)}(\mathbf{x}_s|T_q;\boldsymbol{\varphi})W^{(j)}(T_q)}{\sum_{q=1}^{Q} P^{(j)}(\mathbf{x}_s|T_q;\boldsymbol{\varphi})W^{(j)}(T_q)}\right], \quad (5)$$

where $P^{(j)}(T_q|\mathbf{x}_s)$ is the posterior probability of T_q given response pattern s and is based on item parameter estimates and estimated weights from the previous iteration j (Mislevy & Bock, 1985).

These Q weights, denoted by $\boldsymbol{\lambda} = W(\boldsymbol{T}) = [W(T_1), W(T_2), \ldots, W(T_Q)]$, are estimated with the item parameters in the M-step and then used in the next iteration. Lewis (1985) provided a concise statement about the M-step in the EH characterization of $g(\theta)$:

$$F\left(\boldsymbol{\varphi}^{(j+1)}, \boldsymbol{\lambda}^{(j+1)} \middle| \boldsymbol{\varphi}^{(j)}, \boldsymbol{\lambda}^{(j)}\right) = E\left\{\log f\left(\mathbf{X}, \boldsymbol{\theta} \middle| \boldsymbol{\varphi}^{(j+1)}, \boldsymbol{\lambda}^{(j+1)}\right) \middle| \mathbf{X}, \boldsymbol{\varphi}^{(j)}, \boldsymbol{\lambda}^{(j)}\right\}. \quad (6)$$

Equation 6 gives the quantity that is to be maximized with the EH method. It is apparent from Equation 6 how the item parameters and the distributional parameters are simultaneously estimated. The starting values, $\boldsymbol{\varphi}^{(1)}, \boldsymbol{\lambda}^{(1)}$, are used in the E-step to compute the expected counts, $N_q^{(1)}$ and $R_{iq}^{(1)}$ needed to estimate item parameters. Then, in the M-step, $\boldsymbol{\varphi}^{(1)}$ is updated to $\boldsymbol{\varphi}^{(2)}$ by fitting the logistics using $\boldsymbol{\varphi}^{(1)}$ and $\boldsymbol{\lambda}^{(1)}$. Then, $\boldsymbol{\lambda}^{(1)}$ is updated to $\boldsymbol{\lambda}^{(2)}$, and both $\boldsymbol{\varphi}^{(2)}$ and $\boldsymbol{\lambda}^{(2)}$ are used in the next E-step and so on. There are identifiability concerns when using the EH method for $g(\theta)$ with the 3PL IRT model (Haberman, 2005), as the number of estimated parameters may be very large depending on test length and the number of quadrature points.

Other Approaches to Specify g(θ)

There are a variety of alternative methods to specify a nonnormal latent trait distribution, none of which have gained enough popularity to replace the EH approach as the most widely used. Xu and Jia (2011) used a parametric approach by estimating parameters of the skew-normal distribution in MML (mean, variance, and skewness). Thissen's Johnson curve approach estimates Johnson curve

parameters (Johnson, 1949) along with item parameters in MML yielding curves with different combinations of skewness and kurtosis (available in MULTILOG 7.0 software; Thissen, 2003; van den Oord, 2005).

Another approach to estimating $g(\theta)$ involves Davidian curves (DCs), which are characterized by the product of a squared polynomial of order k and the standard normal density function. To improve estimation and numerical stability, Zhang and Davidian (2001) reparameterized the coefficients of this model using polar coordinates. Woods and Lin (2009) introduced Davidian curve IRT (DC-IRT) with this reparameterized version of DCs. Under this approach, a best fitting solution must be selected from a set of 10 possible DCs, each with a different u value (with $u = 1$ the normal model and higher values of u accommodating various shapes including skewness and multiple modes).

Ramsay-curve IRT (RC-IRT; Woods & Thissen, 2006) implements MML estimation by using a splines-based approach to estimate a smooth $g(\theta)$. That is, in addition to item parameters, RC-IRT estimates parameters which characterize a RC for $g(\theta)$ based on piecewise polynomial "B-spline" basis functions of a specific order and number of breaks (the points where the B-splines join together). The software used to implement RC-IRT provides the user 25 candidate RCs from which to choose the "best" solution based on a comparison of a series of model fit indices from all possible models starting from the simplest 2-2 normal model ("2-2" for knot-order combination), which returns a normal distribution, to the 6-6 model, which is order 6 with 6 knots. In addition to selecting a model based on these two parameters, the user must use trial and error to select the standard deviation (*SD*) for the multivariate normal prior on the spline coefficients. For maximal flexibility in the shape of the distribution, this is typically set to a high value such as 500 and then reduced if there exist estimation issues. Simulation study results show that RC-IRT produces item parameter estimates superior to using a fixed parametric normal distribution, when the true latent trait distribution is nonnormal (Woods & Thissen, 2006). With recent attention on RC-IRT in the literature, we include this approach in our empirical examples. In the next section, we first discuss LLS in terms of observed score distributions and then LLS for the latent trait distribution in the context of unidimensional IRT models.

Loglinear Smoothing

Holland and Thayer (1987, 2000) introduced LLS as a method for reducing irregularities in discrete observed test score distributions. LLS estimates probabilities for discrete distributions based on an unsaturated loglinear model, and therefore, only some properties of the observed frequency distribution are preserved by fitting M moments. This section discusses LLS models for observed and latent distributions.

Model Features and Specification for Observed Distributions

Let θ be a discrete random variable that can take on only the Q values $T_1, T_2, \ldots, T_Q$, with probabilities $p_1, p_2, \ldots, p_Q$, respectively. LLS estimates the probabilities p_q from the contingency table of observations of each value of θ, $T_1, T_2, \ldots, T_Q$, using the polynomial loglinear model

$$\log_e(p_q) = \beta_0 + \sum_{m=1}^{M} \beta_m T_q^m, \tag{7}$$

where T_q^m is the mth power of T_q, and the coefficients $\boldsymbol{\beta}^t = [\beta_1, \beta_2, \ldots, \beta_M]$ and intercept β_0 are to be estimated from the observed counts $n = [n_1, \ldots, n_Q]$. Note that β_0 is a normalizing constant forcing the sum of the p_q to be 1. The same model can be used for the frequencies directly. That is, if p_q satisfies a loglinear model, then n_q will satisfy the same model with β_0 replaced by $\beta_0 + \log N$ (Holland & Thayer, 2000). Therefore, note that $N = \sum n_q$.

The main feature of LLS is that it matches sample moments of the observed distribution. The maximum likelihood estimates from the model in Equation 7, $\hat{\boldsymbol{\beta}}$, force the estimated probabilities to satisfy moment-matching conditions put forth by the model specification (M) and the observed distribution (Holland & Thayer, 2000). The maximum likelihood estimates $\hat{\beta}_m$ satisfy the property that

$$\sum_q T_q^m \hat{p}_q = \sum_q T_q^m (n_q/N), \tag{8}$$

that is, the sample moments of θ match the theoretical moments under the fitted model. The degree of smoothness (or actually "roughness") in LLS is determined by the highest power M of T_q in Equation 7. For $M = 0$ (i.e., not including T_q in the model at all), LLS maximally smooths the $p_q's$, estimating them as a uniform distribution. For M sufficiently large, the loglinear model in Equation 7 is saturated and LLS estimates the p_q as the natural method of moments estimators n_q/N. Indeed, for $M = Q - 1$, the polynomial on the right-hand side of Equation 7 will be an interpolating polynomial for the log $p_q's$. Once the parameters are estimated, the estimated probabilities are computed to characterize the smooth fitted distribution.

LLS for the Latent Trait Distribution

The LLS model for latent distributions is the same as Equation (7); however the values of $T_1, T_2, \ldots, T_Q$ now represent values for quadrature points (or latent trait levels) and the estimates of p_q, quadrature weights, or $W(T_q)$. Under this formulation, there are M parameters estimated for the latent distribution. Since the expected frequencies of test takers at each T_q are unobserved, the EM algorithm includes the LLS procedure to estimate $\boldsymbol{\beta}$.

Consider the number of parameters estimated for $g(\theta)$, and then compare the M moments estimated with LLS to the $Q - 1$ quadrature weights estimated with the EH method. That LLS can result in a substantial reduction in the number of parameters estimated when used instead of the traditional EH method is a fact that cannot be understated. This reduction occurs when $M < Q$. In terms of model parsimony for any test form, LLS can be used with any value of Q and still only M parameters are needed to characterize $g(\theta)$. Estimating $g(\theta)$ with LLS permits the user to estimate nonnormal latent trait distributions without assuming a functional form and without estimating many additional parameters. In addition, LLS allows the user flexibility to control under- and overfitting by manipulating M. In this sense, a M-moment LLS model for $g(\theta)$ provides a "sweet spot" for the bias–variance trade-off, where the levels of bias and variance are optimized, resulting in the least error.

Choosing M *and* Q

The more complex the distribution represented by the $p_q's$, the more moments may be needed. For example, a model that fits two moments can exactly capture a discretized normal distribution. By increasing the number of fitted moments, additional properties of discrete distributions will be captured (e.g., when $M = 3$, skewness is recovered, etc.). In the saturated case, where $M = Q - 1$, the distribution is perfectly fit. For this scenario, the model is not a smoothing model. Univariate observed score distributions typically need four or more fitted moments to be adequately characterized (Holland & Thayer, 2000); however, there are some differences in the number of moments appropriate for fitting a latent trait distribution (compared to an observed distribution). Cressie and Holland (1983) showed that under a Rasch model with both points and weights of $g(\theta)$ estimated, the distribution of the latent trait is determined only up to its first k moments. Specifically, when k, the number of items, is small, then the use of many moments is not supported for latent trait distributions, as the amount of information that can be used in determining probabilities from the distribution is limited.

In terms of choosing Q, generally, the larger the value of Q, the finer the characterized distribution, or the more closely the estimated distribution will approximate a continuous distribution. It may be inferred that the more points specified for the discrete distribution, the better the approximation to $g(\theta)$ and thus better MLEs for the item parameters. Therefore, it is desirable to have a large Q in order to capture the complexities in the distribution, but the benefit may taper. The maximal number of discrete points needed when estimating both points and weights of $g(\theta)$ has been discussed for the Rasch model—it was found that approximately $Q = k/2$ is sufficient (De Leeuw & Verhelst, 1986; Lindsay, Clogg, & Grego, 1991). Tzamourani and Knott (2002) found in tests with k ranging from 3 to 21 that the 2PL model needs fewer than $Q = k/2$. Importantly,

Tzamourani and Knott (2002) also showed in an empirical example that item parameter estimates resulting from varying levels of Q do not differ by much, and discrimination parameter estimates get smaller as Q decreased. They note that this effect on the discrimination parameter estimates is likely a function of coarsening the scale upon which the slope can be appropriately estimated—it is more challenging to discriminate examinees across fewer points or when they are categorized into fewer ability levels.

Note that with LLS, we estimate only M parameters in addition to item parameters, and therefore, the value of Q has no influence on parsimony. In other words, we may choose a relatively larger value of Q to characterize a fine version of $g(\theta)$ and still only need to estimate M parameters.

The EM Algorithm With LLS

To implement LLS for $g(\theta)$ within the MML estimation of item parameters, the loglinear model estimation must be embedded within an EM algorithm (Tsutakawa, 1985; von Davier, 2005; Xu & von Davier, 2008). The E-step is identical for the EH method and LLS, thus details on the E-step may be found elsewhere (Mislevy & Bock, 1985). Therefore, for brevity, we present only the M-step portion of the EM algorithm using LLS for the 3PL model, and the 1PL and 2PL IRT models may also be fitted using variations of this algorithm. We provide the full EM algorithm with LLS in the Appendix.

M(aximization) Step: Stage 1

For each item separately, solve the maximum likelihood equations given in Equations 9, 10, and 11, using the expected counts N_q and R_{iq} computed in the E-step. Note that g_i is the item intercept, a_i is the slope, and c_i is the guessing parameter for item i.

$$\mathrm{g}_i = -\mathrm{a}_i\mathrm{b}_i: \quad 0 = \sum_{q=1}^{Q}(R_{iq} - P_{iq}N_q)H_{iq}, \tag{9}$$

$$\mathrm{a}_i: \quad 0 = \sum_{q=1}^{Q}(R_{iq} - P_{iq}N_q)H_{iq}T_q, \tag{10}$$

$$\mathrm{c}_i: \quad 0 = (1-\mathrm{c}_i)^{-1}\sum_{q=1}^{Q}(R_{iq} - P_{iq}N_q)/P_{iq}. \tag{11}$$

Here, $H_{iq} = \frac{(1-\mathrm{c}_i)P^*_{iq}(1-P^*_{iq})}{P_{iq}(1-P_{iq})}$ and $P^*_{iq} = \psi(\mathrm{a}_iT_q + \mathrm{g}_i)$, the logistic function evaluated at $\mathrm{a}_iT_q + \mathrm{g}_i$. Note that Equations 9, 10, and 11 treat R_{iq} and N_q as known frequencies, but in fact, the posterior expected values are used instead.

M(aximization) Step: Stage 2

Begin a Newton cycle either by using a set of starting values for $\boldsymbol{\beta}^{(1)}$ for the first iteration $v = 1$ in a series of Newton cycles (as described by Holland and Thayer, 1987, pp. 14–15) or by using $\boldsymbol{\beta}^{(v)}$ estimated from the previous cycle if $v > 1$. Note that there may be J iterations of the EM cycles, and within each EM cycle, there may be V iterations of Newton cycles.

(a) Compute a vector of fitted frequencies defined by:

$$\mathbf{f}^{(j)'} = \left[\frac{N^* \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_1^m\right)}{\sum_{q=1}^{Q} \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_q^m\right)}, \frac{N^* \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_2^m\right)}{\sum_{q=1}^{Q} \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_q^m\right)}, \cdots, \frac{N^* \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_Q^m\right)}{\sum_{q=1}^{Q} \exp\left(\sum_{m=1}^{M} \beta_m^{(v)} Z_q^m\right)} \right], \tag{12}$$

where elements of $\mathbf{Z}$ (a $Q \times M$ matrix) are standardized T_q values, taken to the powers 1 through M.

(b) Compute an estimated variance–covariance matrix for this vector defined by $\boldsymbol{\Sigma}_{\mathbf{f}}^{(j)} = \mathbf{D}_{\mathbf{f}}^{(j)} - \frac{1}{N}\mathbf{f}^{(j)}\mathbf{f}^{(j)'}$, where $\mathbf{D}_{\mathbf{f}}^{(j)}$ is a diagonal matrix with the elements of $\mathbf{f}^{(j)}$ in the diagonal.

(c) Using a singular value decomposition, solve for $\boldsymbol{\delta}^{(v)}$ in $(\mathbf{Z}'\boldsymbol{\Sigma}_{\mathbf{f}}\mathbf{Z})\boldsymbol{\delta}^{(v)} = \mathbf{Z}'(\mathbf{n} - \mathbf{f})$, where $\mathbf{Z}$ has columns that are vectors of (standardized) T_q values, taken to the powers 1 through M and $\mathbf{n}' = [N_1, N_2, ..., N_Q]$ is the vector of expected frequencies computed in the E-step.

(d) Compute a new vector of coefficients $\boldsymbol{\beta}^{(v+1)}$ by adding the vector of changes to the vector of coefficients from the previous Newton cycle or the starting values if in the first Newton cycle: $\boldsymbol{\beta}^{(v+1)} = \boldsymbol{\beta}^{(v)} + \boldsymbol{\delta}^{(v)}$.

(e) With the estimates $\boldsymbol{\beta}^{(v+1)}$, and the predefined (and fixed) quadrature point values, compute a vector $\mathbf{f}^{(j+1)}$ of new estimated counts at each quadrature point.

(f) Use the maximum absolute difference between the elements of the original $\mathbf{f}^{(j)}$ and $\mathbf{f}^{(j+1)}$ to determine convergence. If this difference is less than $0.0001 \times N$ (or some other preferred convergence criterion), then convergence for the estimated frequencies has been reached. If convergence has been reached for the LLS algorithm but *not* for item parameters in Stage 1, then continue the EM iterations by returning to the E-step, using $\mathbf{f}^{(j+1)}$ in the next E-step for quadrature weights, such that $W(T_q) = f_q^{(j+1)}/N$. If convergence has not been reached within the Newton cycles, use $\boldsymbol{\beta}^{(v+1)}$ in Step (a) and repeat.

The final output of the M-step is an updated value for the fitted frequencies, which appear as the final quadrature weights, $W(T_q) = f_q/N$.

The software LLSEM 1.0 (*LogLinear Smoothing Expectation Maximization*) implements these algorithms (Casabianca & Lewis, 2011b). LLSEM is capable of item parameter estimation under the 1PL, 2PL, and 3PL dichotomous IRT

models using the normal distribution assumption, the EH method, and LLS. In our implementation of LLS, we keep the quadrature points equally spaced and fixed throughout the entire estimation procedure to identify the scale (Lewis, 1985; see the Appendix for more information on this identification method). Furthermore, there is no standardizing during the iterations, unlike what occurs in many commercial software programs. Instead, we scale item parameter estimates after convergence to the scale of the estimated latent trait distribution. Thus, the mean and *SD* of the resulting distribution are not necessarily 0 and 1, respectively, as would be the result in many commercial software programs. More information on LLSEM is available in the Appendix.

Simulation Study Design

To provide an evaluation of LLS's utility for item parameter recovery, we conducted a Monte Carlo simulation study for IRT item parameter estimation and compared LLS to the fixed normal model and the EH method.[1] We varied skewness and number of modes of the true $g(\theta)$, number of moments fitted, and number of quadrature points. Test length and sample size are popular factors included in item parameter recovery studies; however, they were not included here because pilot study work on LLS for IRT has shown minimal differences with different sample sizes ($n = 500$, 1,000, and 2,000) and test lengths ($k = 25$ and 50; Casabianca, Xu, Jia, & Lewis, 2010). We chose to present results from a simulation study using the 3PL IRT model because that is the model with which there are the most identifiability issues and where parsimony is of the utmost importance. 1PL simulation study results can be found in Casabianca (2011) and the empirical examples found in the following sections use the 2PL model.

Item Response Generation Conditions

Fifty item responses were generated for 1,000 simulated examinees from each of the following true latent trait distributions, each with mean and variance equal to 0 and 1, respectively: (a) standard Normal density, $N(0,1)$; (b) negatively skewed continuous distribution with a skewness of -1.5, created using a mixture of two normal distributions ($\mu_1 = -1.259$, $\sigma_1^2 = 1.791$, $\mu_2 = 0.315$, $\sigma_2^2 = 0.307$, mixing probabilities $m = 0.2$, and $1 - m = 0.8$); and (c) bimodal continuous distribution, also created as a mixture of two normal distributions ($\mu_1 = -0.705$, $\sigma_1^2 = 0.254$, $\mu_2 = 1.058$, $\sigma_2^2 = 0.254$, and $m = 0.6$). One hundred (100) replications of 50,000 item responses (50 items for 1,000 examinees) were generated for each of these three $g(\theta)$ conditions. Figure 1 plots the densities for the three population latent trait distributions. Based on sample estimates from the 100 replications of 1,000 θs, the skewness and kurtosis of the negatively skewed $g(\theta)$ were -1.5 and 3.2, respectively. For the bimodal $g(\theta)$, estimates of these moments were 0.3 and

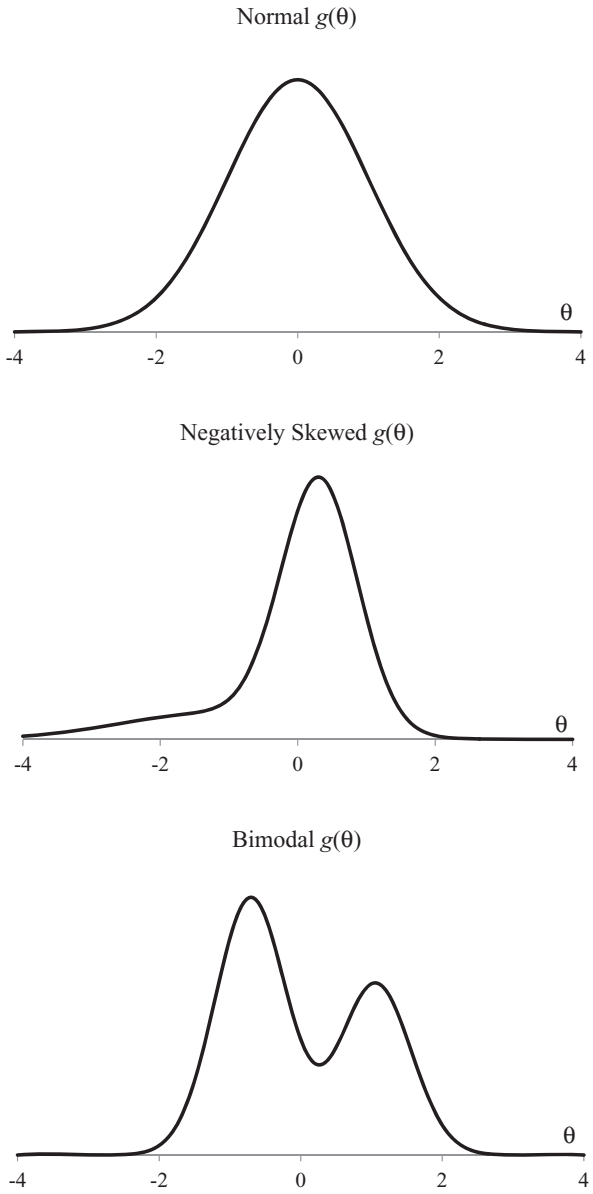

FIGURE 1. *True population latent trait distributions.*

-1.0. The negatively skewed $g(\theta)$ was comparable in terms of the magnitude of skewness to other simulation studies (Casabianca, Xu, Jia, & Lewis, 2010; Woods, 2008; Woods & Lin, 2009; Woods & Thissen, 2006). The mixture

distribution parameters for the bimodal $g(\theta)$ were chosen to be consistent with previous literature that also examined item parameter recovery for a bimodal $g(\theta)$ (Woods & Lin, 2009).

Item responses were simulated using calibrated 3PL items from a 2008 National Mathematics Assessment (Item parameters were provided by ETS; Copyright (C) 2011 ETS. www.ets.org). The distribution of the difficulty parameters was approximately normal (symmetric) but centered around .2. The mean a parameter was 1.13 ($SD = 0.25$), mean b parameter was 0.21 ($SD = 0.51$), and mean c parameter was 0.16 ($SD = 0.05$).

Calibration Conditions

Each data set was calibrated in LLSEM with the following specifications for $g(\theta)$: fixed normal distribution (Mislevy & Bock, 1985), EH method (Mislevy & Bock, 1985), and LLS (Casabianca, Xu, Jia, & Lewis, 2010; Holland & Thayer, 1987, 2000; von Davier, 2005; Xu & von Davier, 2008). There were three levels of the number of quadrature points: $Q = 11$, 31, and 61. Eleven quadrature points were chosen to be close to the default in BILOG-MG, which is 10 (Zimowski et al., 2003). In addition, we chose a higher number not exceeding the test length (31) and a higher number exceeding the test length (61). The last level of Q was selected to support a comparison to other studies that used very large Q (e.g., $Q = 121$; Woods & Lin, 2009; Woods & Thissen, 2006).

The levels of M were specified in a nesting structure according to Q. For all levels of Q, LLS models matching $M = 2$, 3, 4, and 5 moments were fit. These levels were chosen for two reasons. First, what the first few moments can capture can be hypothesized and conceptualized, higher moments probably cannot. Second, based on work with observed test score distributions, it has been noted that at least four or five moments are typically needed to adequately characterize a univariate distribution (Holland & Thayer, 2000). Beyond the five-moment model, M varied with Q in such a way that there was not an excessive number of conditions, but to include enough levels that questions regarding the effect of moment matching on item parameter recovery could be addressed. We aimed to find the optimal point in the bias–variance trade-off. We fitted the 10-moment model for $Q = 31$ and the 10- and 15-moment models for $Q = 61$. Hereinafter, we will refer to LLS models using their M value—for example, LLS5 is the five-moment LLS model. In total, we report results for 63 conditions or 21 conditions per type of true latent trait distribution (normal, negatively skewed, and bimodal).

Outcome Measures

Item parameters are the sole focus of this simulation study. We compared item parameter estimates to the true item parameters using bias and root mean square

error (RMSE) criteria. We computed the average RMSE using the square root of the average mean square error across the 50 items. We also used a multivariate measure of error to assess overall recovery of item parameters called the mxD as used by Woods (2008). Specifically, this measure is the absolute difference between the item characteristic curve (ICC) using estimated item parameters and the ICC using the true item parameters, computed for each item across the Q quadrature points. Within condition, item, and replication, the maximum of these absolute differences (over the Q quadrature points) was determined. The mean of the absolute maximum difference was taken across the 50 items, and the mean was also taken across replications.

Simulation Study Results

Normal g*(θ)*

There was very little error in the normal case to start with! The mxD values for the LLS models (0.044–0.046) were the same as the EH results (0.045–0.046; see Table 1 for these mxD values). Differences between LLS models (across values of M) were negligible, and there was no pattern of increasing or decreasing as a function of the number of moments. There was no difference between $Q = 31$ and $Q = 61$, since virtually all model specifications yielded the same amount of error, on average. However, for $Q = 11$, the average RMSE with the fixed normal distribution specification was 0.01 higher than all other conditions. This difference appeared when examining differences in average RMSEs for the individual item parameters. That is, compared to the $Q = 31$ and $Q = 61$ conditions, the $Q = 11$ condition yielded about 0.03 higher average RMSE for the a and b parameters when using a fixed normal distribution for $g(\theta)$.

Negatively Skewed g*(θ)*

For the negatively skewed $g(\theta)$, mxD was always smallest (for all Q) under LLS4 (0.046–0.047). The EH mxDs were about 0.003–0.007 greater than the optimal LLS4 model.

Figure 2 provides profile plots per parameter estimate and for the true negatively skewed (left column of plots) and bimodal (right column of plots) latent trait distributions. These plots show the trajectory for error for the three different Q levels, as we increase in model complexity. Here, $\times$ is $Q = 11$, $\bigcirc$ is $Q = 31$, and Δ is $Q = 61$. We excluded the profile plots for the true normal $g(\theta)$ because there were very few differences between models. Of the LLS models, Figure 2 shows that the average RMSEs were largest for LLS2, decrease from LLS3 and LLS4, and then showed either a slight increasing trend or tapered off as M increased. The a parameter was recovered best with LLS4 with $Q = 11$. The lowest average RMSEs were found with LLS4 for $Q = 11$ and the LLS5 for $Q = 31$ and 61. For the b parameters, LLS4 was the best model (for all values of Q).

TABLE 1.
Maximum Absolute Difference (mxD) in Item Characteristic Curves (Estimated vs. True)

Normal $g(\theta)$					Negatively Skewed $g(\theta)$					Bimodal $g(\theta)$				
Method	M	$Q = 11$	$Q = 31$	$Q = 61$	Method	M	$Q = 11$	$Q = 31$	$Q = 61$	Method	M	$Q = 11$	$Q = 31$	$Q = 61$
Normal		.054	.043	.043	Normal		.089	.079	.079	Normal		.055	.058	.058
LLS	2	.044	.045	.045	LLS	2	.077	.081	.081	LLS	2	.064	.067	.067
LLS	3	.046	.046	.046	LLS	3	.047	.052	.051	LLS	3	.078	.071	.072
LLS	4	.045	.046	.046	LLS	4	.046	.047	.047	LLS	4	.045	.050	.050
LLS	5	.045	.046	.046	LLS	5	.049	.048	.049	LLS	5	.048	.053	.053
LLS	10		.046	.046	LLS	10		.050	.050	LLS	10	.048	.052	.052
LLS	15			.046	LLS	15			.050	LLS	15			.053
EH		.045	.046	.046	EH		.053	.050	.050	EH		.048	.053	.053

Note. Normal = fixed (standard) normal distribution assumption; LLS = loglinear smoothing; EH = empirical histogram method.

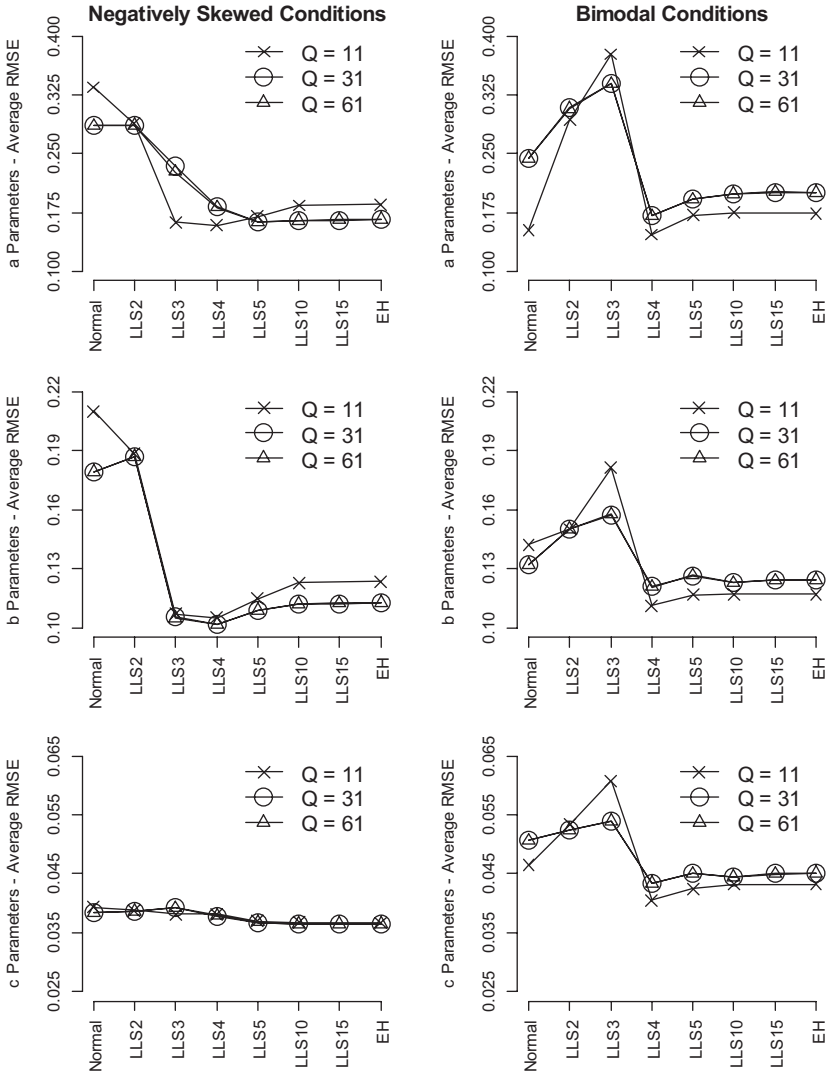

FIGURE 2. *Average root mean square error plots for* a, b*, and* c *item parameter estimates for negatively skewed and bimodal* g(*θ*) *conditions. The left column of plots is for the negatively skewed conditions and the right column for the bimodal conditions. LLSM = M-moment LLS model; EH = empirical histogram; Normal = fixed normal distribution assumption.*

Finally, there was no trend for the *c* parameter. Figure 2 shows no distinction between specifications, although there was a slight dip in error in LLS10 and LLS15.

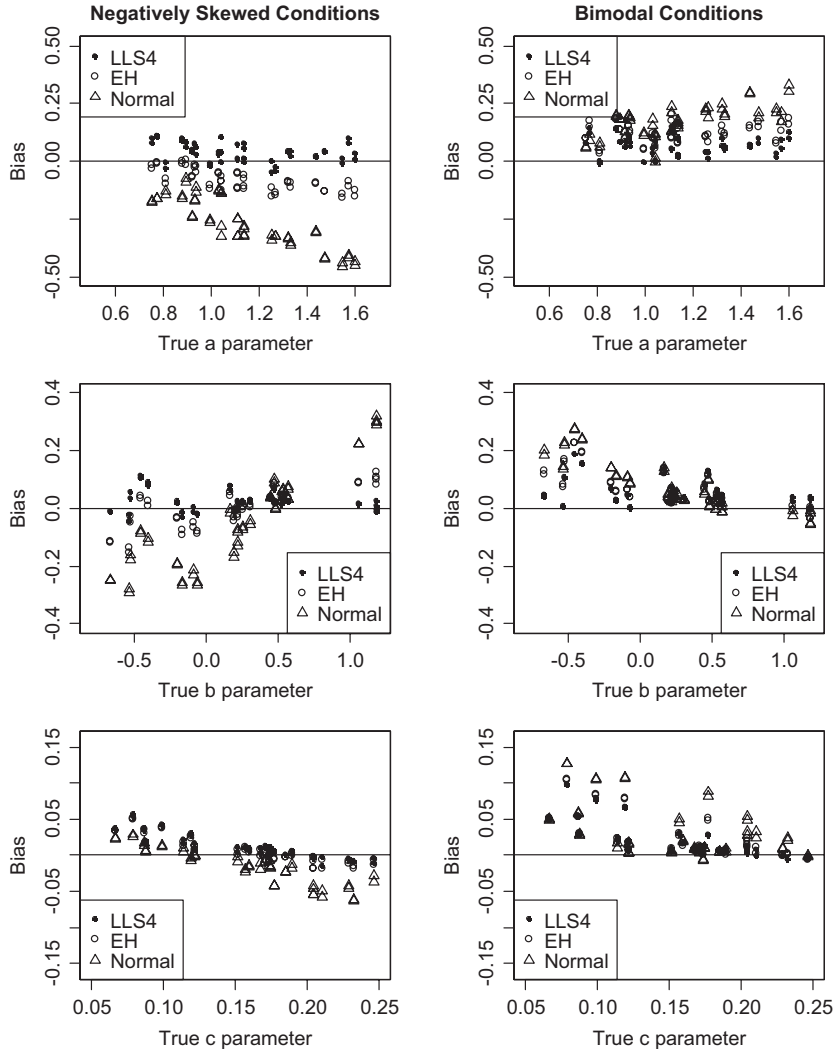

FIGURE 3. *Item bias plotted as a function of true item parameters for 50 items in simulation study under the normal, EH method, and four-moment loglinear smoothing (LLS) models when* Q = *31. The top, middle, and bottom rows show the bias for the* a, b, *and* c *parameters, respectively. The left column of plots is for the negatively skewed conditions and the right column for the bimodal conditions. LLS4 = four-moment LLS model; EH = empirical histogram; Normal = fixed normal distribution assumption.*

There was no difference between $Q = 31$ and 61. However, when $Q = 11$, the fixed normal distribution assumption yielded a mxD of 0.089, which is higher than mxD = 0.046 yielded by LLS4. This difference is reflected in Figure 2 in the *a* and *b* parameters. Under LLS3, the $Q = 31/61$ conditions yielded larger average RMSEs than $Q = 11$ for *a* parameters. For both the *a* and *b* parameters, the average RMSEs were slightly larger for the $Q = 11$ condition under LLS5.

The left column of plots in Figure 3 depict item-level bias for the negatively skewed condition for the *a*, *b*, and *c* parameters (going down the column) under the fixed normal assumption, EH, and LLS4 with $Q = 31$. For brevity, only LLS4 was shown here (in filled circles) because it has been shown throughout the results to be the top performer. LLS4 yielded consistent positive bias across the *a* parameter scale with fluctuations around 0. EH and especially the fixed normal model showed a downward bias as *a* increased. Bias in the *b* parameters was usually smaller with LLS4 than the other models—this was especially true at the ends of the scale. In a couple of instances, the LLS4 model yielded approximately 0 bias, while EH yielded about 0.1 (in both the positive and negative directions). All three models had similar levels and trends for bias for the *c* parameter: positive bias for lower *c* parameters and negative bias for higher *c* parameters. LLS4 had relatively smaller biases with higher *c* parameters, although the highest true *c* in this simulation was only .25.

Bimodal g(θ)

For the bimodal condition, the mxD ranged from 0.045 to 0.078 when $Q = 11$ and 0.050 to 0.072 when $Q = 31/61$. The absolute least error was obtained with $Q = 11$ under LLS4 (mxD = 0.045). All LLS models performed similar to EH except for LLS2 and LLS3, which yielded more error. Although differences between the models were minimal, the mxD and the average RMSEs for LLS4 were always smaller (mxD difference = 0.003) than the EH results.

All three types of item parameters also showed the smallest average RMSEs under LLS4 (see Figure 2). There was an increase in error between LLS2 and LLS3 and then a large decrease in error between LLS3 and LLS4. After $M = 4$, the mxD and the average RMSEs increased, but only slightly.

Differences between levels of *Q* were not straightforward. The two higher levels of *Q* were identical and there were inconsistent differences between $Q = 11$ and $Q = 31/61$. For example, for all models but the LLS3 model, mxD was smaller when $Q = 11$. Conversely, for LLS3, mxD was larger when $Q = 11$. This is also shown on the profile plots (Figure 2) for each of the item parameters.

The right column of Figure 3 shows item-level bias plots for the bimodal condition. In the top plot, the bias for the *a* parameters increases with the true *a* value for the fixed normal distribution assumption and EH, but this trend was not true for LLS4. Instead, bias tended to be smaller under LLS4. For the lower *b*

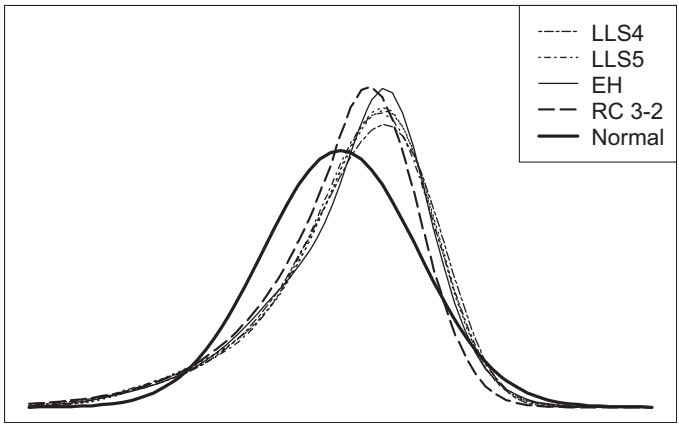

FIGURE 4. *Estimated latent trait distributions with Q = 61 from Programme for International Student Assessment mathematics data (Shanghai-China sample). LLS4 = four-moment LLS model; LLS5 = five-moment loglinear smoothing (LLS) model; EH = empirical histogram method; RC 3-2 = 3-2 Ramsay-curve item response theory solution; Normal = fixed normal distribution assumption.*

parameters (between −0.7 and 0), the LLS4 model yielded smaller (positive) bias than the other models. Toward the center of the *b* scale, the three models converged. Larger positive bias was found for the smaller *c* parameters (0.07 to 0.18), and LLS4 yielded the smallest of these biases. As *c* increased, bias for all models converged around 0.

Empirical Illustration Using the PISA Mathematics Assessment—Shanghai Sample

The PISA is an international study conducted by the Organization for Economic Cooperation and Development that assesses 15-year-old students in mathematics, reading, and science. Shanghai-China outperformed all other countries in all three PISA subject assessments in 2012. We analyzed 11 dichotomously scored mathematics items from a subset ($n = 1{,}623$) of the Shanghai-China sample ($n = 5{,}177$) to demonstrate the 2PL EM algorithm with LLS for $g(\theta)$ and compare it to other methods.[2] Four calibration methods (fixed normal assumption, EH, LLS, and RC-IRT) were compared using LLSEM 1.0 and RCLOG v.2 (Woods, 2006b). We set the bounds of the distribution to (−4, 4), used 61 quadrature points, and fitted LLS models with $M = 4$ and 5.

In our RCLOG specifications, we set the *SD* for the multivariate normal prior on the spline coefficients to 500 for maximal flexibility in the shape of the

distribution. We fitted all possible RC-IRT models with order up to 6 and number of breaks up to 6 and selected the best fitting model based on the consideration of several criteria as described in Woods (2006a). The criteria are the log-likelihood LogL, the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the Hannan–Quinn criterion (HQ; Hannan, 1987), and the Kolmogorov–Smirnov test for normality (KS; Kolmogorov, 1933; Smirnov, 1939). The models with the smallest values of the LogL, AIC, BIC, and HQ signal the best fit. Woods (2006a) notes that it is common for these various criteria to yield different models; however, the current analysis revealed consistency across three of these four indices. That is, the 3-2 model was selected by BIC, HQ, and AIC. The LogL selected the 6-6 model. Both of these models are significantly different from normal model by the KS test. We proceed with the comparison of models using results from 3-2 RC-IRT model solution.

Figure 4 provides the latent trait distributions from the MML estimation of item parameters for the PISA items.[3] The thicker solid curve is the discretized normal distribution. The thinner curves are distributions from EH, LLS, and RC-IRT, all of which should capture varying degrees of nonnormality, if any nonnormality should exist. Indeed, the estimated latent distributions all exhibit moderate positive kurtosis and moderate negative skewness. (For reference, consider statistics for the 3-2 RC: $M = 0.000$, $SD = 1.001$, skewness $= -.884$, kurtosis $= 4.004$.) While they all share similar shapes, the RC 3-2 and EH distributions are more leptokurtic than the LLS distributions. These distributions indicate a population of examinees that tend to perform relatively higher on the latent trait scale, which is expected from the Shanghai-China sample.

The profile plots of the *a* and *b* item parameter estimates from the PISA analysis in Figure 5 provide a visual depiction of the differences between estimates by method for g(θ). The x-axes on these plots arrange items sorted ascending by the normal parameter estimate. When examining the top plot of Figure 5, we see that all methods with the exception of the normal model yield very similar *a* parameter estimates. Only with the most discriminating item (#5) was there a difference between RC-IRT and the other methods. In this case, RC-IRT yielded a marginally higher estimate (2.28 vs. $\sim$2.16 for the other methods). Note that although we observed a slight difference in the distributions for EH/RC-IRT and LLS solutions, the item parameters here are basically equivalent. There were inconsistent differences between the normal model estimates (black line and diamond symbol) and the other methods, in that for some items, the normal estimate was higher and for other items lower. This may be related to the corresponding difficulty estimates for these items (see the bottom plot of Figure 5). For easy items, the normal model yielded larger *b* estimates compared to the other methods.

Although the LLS4 and EH models were very similar in parameter recovery as per the simulation results, it was technically the LLS4 model that yielded the most accurate ICC, when g(θ) was nonnormal (shown via the MxD statistic in

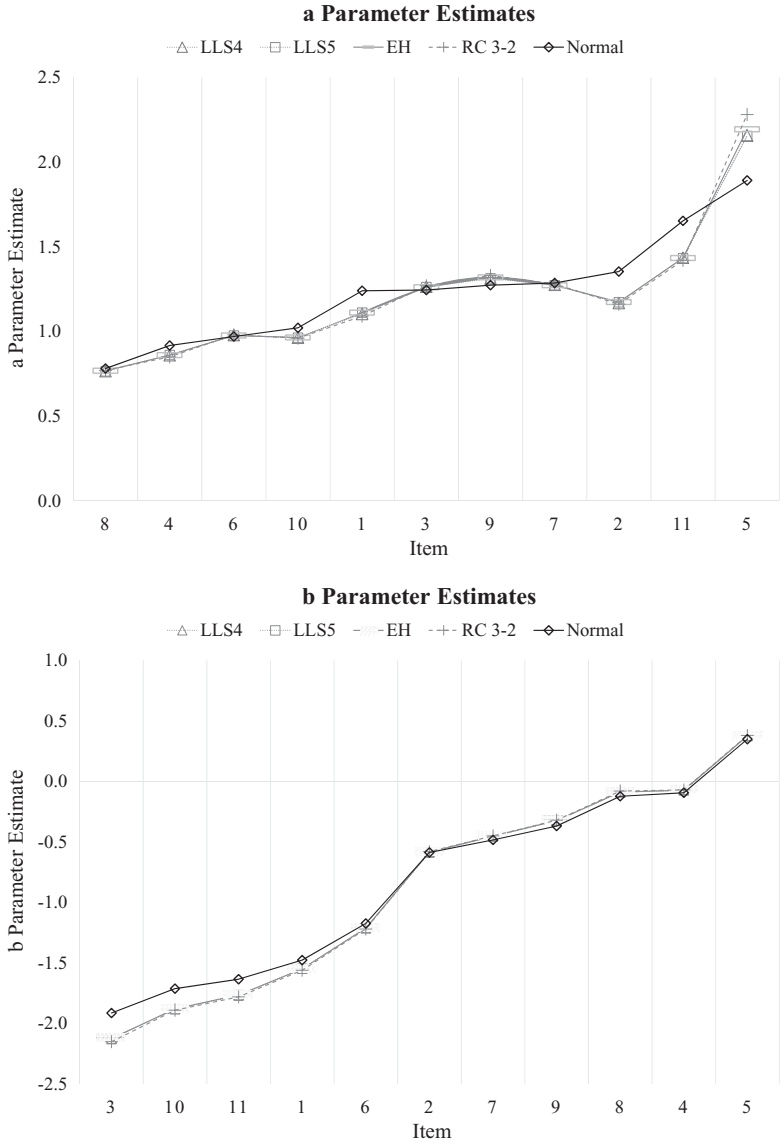

FIGURE 5. *Item response theory (IRT) parameter estimates plotted in profiles by method for g(θ) for 11 mathematics items from the 2012 Programme for International Student Assessment, Shanghai-China sample. The* a *and* b *estimates appear in the top and bottom plots, respectively. Items are sorted in ascending order by the normal parameter estimate. LLS4 = four-moment loglinear smoothing (LLS) model; LLS5 = five-moment LLS model; EH = empirical histogram method; RC 3-2 = 3-2 Ramsay-curve IRT solution; Normal = fixed normal distribution assumption.*

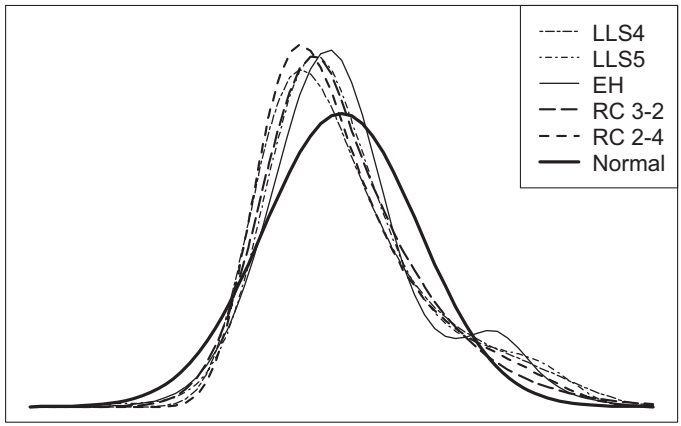

FIGURE 6. *Estimated latent trait distributions with Q = 61 from Maudsley Obsessional Compulsive Inventory wash item data. LLS4 = four-moment LLS model; LLS5 = five-moment LLS model; EH = empirical histogram method; RC 3-2 = 3-2 Ramsay-curve item response theory (IRT) solution; RC 2-4 = 2-4 Ramsay-curve IRT solution; Normal = fixed normal distribution assumption.*

Table 1). The EH MxDs were always larger but by a very small order. It is reasonable to cautiously consider the LLS4 estimates here the best of the normal, EH, and LLS models. However, we must remind you that the distributions in the simulations and in this example differ, and the IRT models differ, making a direct correspondence between this example and the simulations less than straightforward.

Empirical Illustration Using the MOCI

We used data from an administration of the MOCI (Hodgson & Rachman, 1977) to a sample of undergraduates enrolled in an introductory psychology class at the University of North Carolina ($n = 1{,}080$) to provide an additional demonstration of LLS. The MOCI is a multidimensional true/false measure of obsessive–compulsive symptoms. In this illustration, we use 9 items from the MOCI which together are unidimensional and comprise the "wash" subscale (Woods, 2002). The subscale includes items such as "My hands do not feel dirty after touching money." True responses to these items indicate a lack of obsession/compulsion with washing, and therefore, respondents with negative latent trait values have higher levels of obsession and compulsion and the opposite is true as θ increases. We calibrated the items with the 2PL IRT model under the normal, EH, LLS4, LLS5, and RC-IRT with $Q = 61$ and the bounds of the distribution set to $(-4, 4)$. We fitted all possible RC-IRT models yielding 25 candidate RCs using a *SD* equal to 500 for the multivariate normal prior on the spline

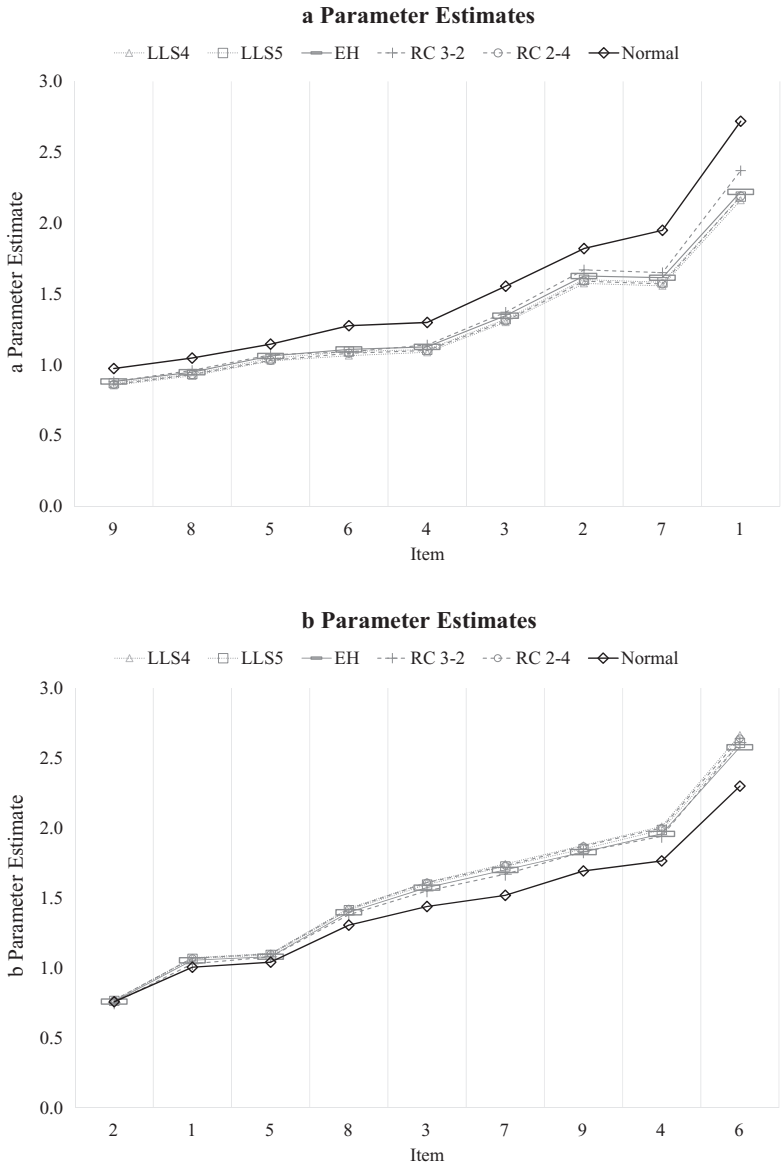

FIGURE 7. *Item response theory (IRT) parameter estimates plotted in profiles by method for* g*(θ) for 9 wash items from Maudsley Obsessional Compulsive Inventory. The* a *and* b *estimates appear in the top and bottom plots, respectively. Items are sorted in ascending order by normal parameter estimate. LLS4 = four-moment LLS model; LLS5 = five-moment LLS model; EH = empirical histogram method; RC 3-2 = 3-2 Ramsay-curve IRT solution; RC 2-4 = 2-4 Ramsay-curve IRT solution; Normal = fixed normal distribution assumption.*

coefficients. The best fitting model according to the BIC and HQ was the 3-2 model. The AIC selected the 2-4 model and LogL selected the 4-3 model. We continued with the 3-2 and 2-4 models in our analysis because the AIC, HQ, and BIC are preferred, as they are penalized likelihood criteria.

Figure 6 shows the series of distributions all with some degree of positive skew and kurtosis (3-2 RC: $M = 0.000$, $SD = 1.002$, skewness $= 0.729$, kurtosis $= 3.801$). The deviations from normality are clear when comparing these to the normal distribution. Interestingly, the EH distribution exhibits a very small secondary mode on the right side—the other methods did not reveal this. It is true that at some point, LLS would also show this same mode, but it is unknown with what value of M. Aside from that difference found in the EH distribution, all other estimated distributions share the same shape.

Figure 7 reveals only small differences between item parameters yielded by the models estimating $g(\theta)$. The RC 3-2 and EH a estimates were almost always larger than as from LLS and RC 2-4. The normal a estimates were always larger, and substantially. This difference between the normal model and the other methods increased as the a value increased. The b parameter estimates based on the normal model ranged from 0.75 to 2.3. There were basically no differences between b estimates for the easier items; however, the normal estimates were always smallest and RC 3-2 bs were also always just slightly smaller than LLS and EH bs. As items become more difficult, the difference between the normal bs and the other bs increased. Slight differences between the methods estimating $g(\theta)$ also appear as b increases.

Differences between the normal model item parameter estimates and estimates from the other methods for $g(\theta)$ are consistent with the item parameter recovery results found in the simulations. We know that the normal model does not perform well when $g(\theta)$ is nonnormal. Therefore, the differences observed here point toward using a nonparametric or semiparametric approach for $g(\theta)$.

Discussion

Our goals were to provide the LLS/EM algorithm for IRT item parameter estimation, evaluate its effectiveness when $g(\theta)$ is nonnormal and with different values for M and Q, and compare it to other methods. This research investigated IRT item parameter recovery using various methods for specifying $g(\theta)$ when the true $g(\theta)$ was normal, negatively skewed, and bimodal. For the negatively skewed and bimodal $g(\theta)$ conditions, LLS models matching four or more moments slightly outperformed or matched the performance of the EH method. In most cases, the differences in these average errors were small; however, we observed substantial differences in bias at the item level. For example, Figure 3 shows that when $g(\theta)$ was negatively skewed, an item with true $b = 1.1$ had 0 bias under LLS4 but bias of 0.1 under EH. Similar differences are observable in Figure 3 for both nonnormal $g(\theta)$ conditions. Differences of this order are substantial,

especially considering the subsequent uses of parameter estimates (e.g., test score equating) and the need for the values entered into these analyses to be precise.

It was generally true that somewhere along the M scale, the amount of error from item parameter estimation came to a stable minimum. In other words, there was a point at which additional moments did not contribute to the characterization of $g(\theta)$ and thus did not contribute to the estimation of item parameters. Furthermore, in some instances, additional moments actually led to poorer quality item parameter estimates, most likely due to overfitting. These asymptotic points occurred at different levels of M for each true $g(\theta)$. For the normal $g(\theta)$, there was a very small dip in error at the two-moment mark, and from the three-moment model on, error was consistently larger (though not by much). However, for both nonnormal distributions, this minimum certainly did not occur at two moments! For the particular negatively skewed $g(\theta)$ examined in this study (skewness $= -1.5$), the four-moment model yielded the least error. Similarly, for the bimodal $g(\theta)$ modeled in this study, the two- and three-moment models yielded more error, and then error dipped to a minimum at four moments. After four moments, increasing the number of moments produced a gradually increasing trend in error. The dip in error represents the optimal point, where bias and variance together lead to the least error. These optimal points are inherently specific to the distributions modeled in this study.

While there was virtually no difference in item parameter recovery for $Q = 31$ and $Q = 61$, there were some effects to the discrimination parameter between $Q = 11$ and the higher Q conditions. Using more points to represent the distribution (keeping the range of the distribution fixed) should impact the estimation of the slope. That is, more points characterize a finer $g(\theta)$, and keeping all else constant, estimates of the slope should increase as Q increases (see Casabianca, 2011 or Tzamourani & Knott, 2002). This was in fact true with the normal $g(\theta)$ condition. There was negative bias when $Q = 11$, which disappeared when $Q = 31/61$. Finally, the fixed normal distribution assumption yielded differences in error between levels of Q. Specifically, there was a difference between the higher levels, $Q = 31/61$ and $Q = 11$. With the fixed normal distribution assumption, error was larger in $Q = 11$ for the negatively skewed condition and error was smaller in $Q = 11$ for the bimodal condition. Inconsistencies in the results are likely attributable to the conflation between the true $g(\theta)$, Q, and M and their interactions. Additional factors including the range of the distribution and the spacing between quadrature points would also affect recovery of the latent trait distribution and item parameters.

The data examples provide a comparison of item parameters and latent distributions generated by a series of methods including RC-IRT for two data sets which both involved distributions with moderate skewness and kurtosis. The simulation and real data results were consistent: with the exception of the normal

model, differences between models for $g(\theta)$ were mostly minor. We did observe many similarities between LLS and EH in both data sets, but we know from simulations that smoothing results in more accurate item parameter estimates.

LLS Versus Other Methods for g(θ)

It is obvious why any nonparametric or semiparametric approach for $g(\theta)$ may be preferred over the fixed normal model—these models will properly model deviations from normality in the characterization of $g(\theta)$ thereby yielding more precise item parameter estimates. The problematic part about the flexibility of the most popular nonparametric approach, the EH approach, is that it involves an estimation procedure requiring possibly many additional parameters. For long tests and for complex models, where many item parameters are being estimated, this may lead to estimation and identifiability issues (Haberman, 2005). LLS offers a unique advantage over the EH method by estimating only the M moments necessary to capture nonnormality, making the LLS solution more parsimonious. We showed in our simulations that in addition to improvements in parsimony, LLS actually yields *better* item parameter estimates in terms of RMSEs and MxDs. However, we must remind the reader that this result is strictly applicable to the data under study.

The payoff for using LLS is potentially great. For example, suppose we have a 50-item test to be calibrated with the 3PL with $Q = 61$. Under the normal model, only 150 parameters are estimated. However, under EH, we estimate $150 + 60 = 210$ parameters. With a five-moment LLS model, we can obtain the same or better item parameter estimates with $150 + 5 = 155$ parameters estimated. The number of additional parameters involved with RC-IRT is equal to the sum of the order and number of breaks minus two. Therefore, for the 3-2 model that was selected for our two empirical examples, we estimated an additional $3 + 2 - 2 = 3$ spline coefficients. For the most complex RC-IRT model, we would estimate $6 + 6 - 2 = 10$ spline coefficients.

RC-IRT is a very flexible option. The RCs estimated and selected for analysis in our empirical examples appear almost identical to the curves from the EH and LLS methods and the resulting item parameter estimates were very similar to these methods as well. The major difference between RC-IRT and LLS is in the implementation. With RC-IRT, the analyst considers a variety of candidate RC curves based on spline functions with different combinations of order and number of breaks and selects a model based on a series of model fit indices. There is also a decision to be made in regard to the *SD* of the multivariate prior distribution. Consequently, there are many possible solutions depending on the final *SD* chosen as well as the model selected based on the multiple fit indices, which often select different models. In our opinion, the wide range of options makes RC-IRT less straightforward than LLS.

Further, because it is a splines-based approach, RC-IRT has the potential for indeterminacy. The estimation procedure depends on the number and location of breaks in the RC used to characterize $g(\theta)$. There is an upper limit on the number of breaks available in RCLOG (2–6 breaks). This dependency may result in estimation issues if there are insufficient breaks or if the breaks appear in inopportune locations. For example, Woods and Thissen (2006) reported issues with convergence because the estimated curve required more quadrature range than was available.

LLS provides smoothed solutions based on a moment-matching procedure that relaxes the potentially overfitted $g(\theta)$ generated by EH. It does this all while yielding the same or better item parameters with less of an expense in terms of the number of estimated parameters. The EH is a special case of the LLS framework (when $M = Q - 1$). The mathematics/statistics underlying LLS is accessible to researchers and analysts and the approach to implementing LLS is straightforward with minimal consideration of anything but moments. Considering that as little as four moments can recover adequately a bimodal distribution, using LLS is very simple with minimal decisions to be made that would otherwise generate inconsistencies in model selection. LLS would not present many indeterminacy issues like RC-IRT because it can easily be implemented with very large Q values to characterize $g(\theta)$.

Suggestions for Use and Final Remarks

One model that kept reappearing as the best in terms of yielding smaller average RMSEs is the four-moment LLS model. Clearly, the two- and three-moment models are generally insufficient. We suggest the four- or even the five-moment models for the future use of LLS to estimate $g(\theta)$ in the 3PL unidimensional IRT context. A more exhaustive approach to model selection would be to first calibrate the items with EH and observe the recovered latent trait distribution. The item responses could then be recalibrated with LLS using an M selected based on the estimated distribution in the first calibration.

There are some caveats that should be noted in the interpretation of the simulation results. First, there is limited generalizability because we explored only two distinct nonnormal latent trait distributions. Further, the different item parameters in the 3PL IRT model will have different levels of robustness to poor characterizations of the latent trait distribution. For example, if the scale is not wide enough (range restriction) or if there is improper representation of the distribution, then there may be problems estimating the a parameter. Also, it is known that there are issues estimating c parameters, and the estimation of the a and c parameters are interdependent as the c defines the lower asymptote for the ICC. Future research will implement LLS in other contexts such as multiple group $g(\theta)$ estimation, where the characterization of the distributions is the primary outcome. The benefits of using LLS may be more apparent when using LLS

in more complex models, such as multidimensional IRT models, where additional research with nonnormal latent variable distributions is needed (Cai, 2010).

Acknowledgment

The authors are extremely grateful to Brian Junker and three anonymous reviewers for their careful reading and useful feedback on this research and article. We are also grateful to Carol Woods for providing the MOCI data and for assisting us with the RCLOG software.

Authors' Note

A version of this article was presented at the 2012 National Council on Measurement in Education annual meeting.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This research is based on the first author's dissertation carried out at Fordham University and was jointly supported by Fordham University's Distinguished Dissertation Research Fellowship (2009) and by Educational Testing Service's (ETS) Harold Gulliksen Psychometric Research Fellowship (2009). The research reported here was also supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B1000012 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education, Fordham University, or ETS.

Notes

1. Unfortunately, we were unable to include Davidian curve-item response theory in our simulations and empirical examples because the software is not currently functional. Our empirical data examples compare Ramsay-curve item response theory (RC-IRT) to loglinear smoothing; however, this method is excluded from the simulation study, as the software to implement RC-IRT cannot be batched processed and the source code is unavailable.
2. We selected a subset of the Shanghai-China sample that was administered booklet numbers 3, 7, 9, and 10, yielding responses to 11 dichotomous items.
3. To facilitate the comparison of distributions, we rescaled the quadrature points for the *LLSEM*–based distributions (LLS4, LLS5, empirical histogram [EH], and Normal) to ensure a mean of 0 and variance of 1.0. However, because each latent distribution has a different mean and variance at the end of the estimation procedure, the points do not perfectly align. This may be resolved by interpolation.

Supplemental Material

The online data supplements are available at http://jeb.sagepub.com/supplemental

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Academiai Kiado.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, *42*, 357–374.

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

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

Boulet, J. R. (1996). *The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods (IRT, nonlinear factor analysis)*. Dissertation abstracts online, University of Ottawa (Canada).

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Casabianca, J. M. (2011). *Loglinear smoothing for the latent trait distribution: A two-tiered evaluation*. (Doctoral dissertation). ProQuest Dissertations and Theses. (Accession Order No. AAT 3474125).

Casabianca, J. M., & Lewis, C. (2011a, April). *The Impact of IRT Item Parameter Estimation Error on IRT True Score Equating Functions and CSEMs: A Robustness Study*. Paper presented at the annual meeting of the National Council for Measurement in Education. New Orleans, Louisiana.

Casabianca, J .M., & Lewis, C. (2011b). *LLSEM 1.0: LogLinear Smoothing in an Expectation Maximization algorithm for item response theory item parameter estimation*. [Computer software]. Bronx, NY.

Casabianca, J. M., Xu, X., Jia, Y., & Lewis, C. (2010, May). *Estimation of Item Parameters when the Underlying Latent Trait Distribution of Test Takers is Nonnormal*. Paper presented at the annual meeting of the National Council for Measurement in Education. Denver, Colorado.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*, 129–141.

De Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*, 183–196.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Eysenck, H. J., & Eysenck, S. B. G. (1991). *Eysenck Personality Questionnaire—Revised (EPQ-R)*. London, England: Hodder and Stoughton.

Haberman, S. J. (2005). *Identifiability of parameters in item response models with unconstrained ability distributions* (Research Report 05-24). Princeton, NJ: Educational Testing Service.

Hannan, E. J. (1987). Rational transfer function approximation. *Statistical Science*, *2*, 135–151.

Hodgson, R. J., & Rachman, S. (1977). Obsessional-compulsive complaints. *Behaviour Research and Therapy*, *15*, 389–395.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of loglinear models for fitting discrete probability distributions* (Research Report 87-31). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*, 149–176.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distributione. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 1–11.

Lewis, C. (1985). Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1982 IRT and Computerized Adaptive Testing Conference* (pp. 203–209). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96–107.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517–549.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Matthews, G., Deary, I. J., & Whiteman, M. C. (2003). *Personality traits* (2nd ed.). New York, NY: Cambridge University Press.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM Algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 IRT and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Lincolnwood, IL: Scientific Software International.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika*, *48*, 567–574.

Rost, J., & von Davier, M. (1992). *MIRA—A PC program for the mixed Rasch model user manual*. Kiel, Germany: IPN.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar. (Eds.), *Rasch models: Foundations, recent developments, and applications*. (Ch. 14; pp. 257–268). New York, NY: Springer Verlag.

Sanathanan, L., & Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, *73*, 794–799.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Smirnov, N. (1939). Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique*, *6*, 3–26.

Stone, C. A. (1992). Recovery of MML estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1–16.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice Hall.

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York, NY: Academic Press.

Thissen, D. (2003). *MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory* [Computer software]. Chicago, IL: Scientific Software.

Tsutakawa, R. K. (1985). Estimation of item parameters and the GEM algorithm. In D. J. Weiss (Ed.), *Proceedings of the 1982 IRT and computerized adaptive testing conference* (pp. 180–188). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, *55*, 371–390.

Tzamourani, P., & Knott, M. (2002). Fully semiparametric estimation of the two-parameter latent trait model for binary data. In G. A. Marcoulides & I. Moustak (Eds.), *Latent variable and latent structure models* (pp. 63–84). Mahwah, NJ: Erlbaum.

van den Oord, E. J. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement*, *29*(1), 45–64.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

Woods, C. M. (2002). Factor analysis of scales composed of binary items: Illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioral Assessment*, *24*, 215–223.

Woods, C. M. (2006a). Ramsay-curve item response theory to detect and correct for non-normal latent variables. *Psychological Methods*, *11*, 253–270.

Woods, C. M. (2006b). *RCLOG v.2: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* Technical Report. St. Louis, MO: Washington University in St. Louis.

Woods, C. M. (2008). Ramsay-curve item response theory for the three-parameter logistic item response model. *Applied Psychological Measurement*, *32*, 447–465.

Woods, C. M., & Lin, N. (2009). IRT with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*, 102–117.

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

Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (Research Report 11-40). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (Research Report 08-27). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Muraki, E. (1991, April). *Non-linear transformation of IRT scale to account for the effect of nonnormal ability distribution on the item parameter*

estimation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*, 795–802.

Zhang, J. (2011, April). *The role of standard errors of item parameter estimators in IRT-based data analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zhang, J., Xie, M., Song, X., & Lu, T. (2011). Investigating the impact of uncertainty about item parameters on ability estimation. *Psychometrika*, *76*, 97–118.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Lincolnwood, IL: Scientific Software International.

Authors

JODI M. CASABIANCA is an assistant professor in the University of Texas at Austin, 1 University Station D5800, Austin, TX, 78712; e-mail: jcasabianca@austin.utexas.edu. Her research interests are educational measurement, psychometrics, and statistics.

CHARLES LEWIS is Professor Emeritus in Fordham University, 441 E. Fordham Road, Bronx, NY 10458; e-mail: clewis@fordham.edu. His research interests include fairness and validity in educational testing, mental test theory, including item response theory and computerized adaptive testing, general(ized) linear models, Bayesian inference, and behavioral decision making.

Manuscript received March 3, 2015
Revision received July 7, 2015
Accepted July 19, 2015