Article

Journal of Educational and Behavioral Statistics
2021, Vol. 46, No. 6, pp. 651–681
DOI: 10.3102/1076998620983908
Article reuse guidelines: sagepub.com/journals-permissions

Cross-Classified Random Effects Modeling for Moderated Item Calibration

Seungwon Chung
University of Minnesota

Li Cai
University of California, Los Angeles

In the research reported here, we propose a new method for scale alignment and test scoring in the context of supporting students with disabilities. In educational assessment, students from these special populations take modified tests because of a demonstrated disability that requires more assistance than standard testing accommodation. Updated federal education legislation and guidance require that these students be assessed and included in state education accountability systems, and their achievement reported with respect to the same rigorous content and achievement standards that the state adopted. Routine item calibration and linking methods are not feasible because the size of these special populations tends to be small. We develop a unified cross-classified random effects model that utilizes item response data from the general population as well as judge-provided data from subject matter experts in order to obtain revised item parameter estimates for use in scoring modified tests. We extend the Metropolis–Hastings Robbins–Monro algorithm to estimate the parameters of this model. The proposed method is applied to Braille test forms in a large operational multistate English language proficiency assessment program. Our work not only allows a broader range of modifications that is routinely considered in large-scale educational assessments but also directly incorporates the input from subject matter experts who work directly with the students needing support. Their structured and informed feedback deserves more attention from the psychometric community.

Keywords: *test modification; expert judgment; crossed random effects; small sample; estimation*

1. Introduction

Test forms are often *modified* to accommodate various special populations or situations where administration of the original test forms is infeasible. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) defines

modification as "changes made in the content, format, and/or administration procedure of a test in order to accommodate test takers who are unable to take the test under standard conditions" (p. 183). It is distinguished from *accommodation* in that it refers to changes that could potentially alter the underlying construct and ultimately compromise the comparability of scores (AERA et al., 2014; Thurlow et al., 2001).

Under Title I of the Every Student Succeeds Act, state education agencies must include all students, including students with disabilities who are English learners (ELs), in their assessment and accountability systems. This reaffirmation of the civil rights of students with disabilities, however, generates a formidable psychometric challenge. For the sake of fairness, scores derived from the modified items should be made comparable to those from the original items developed for the general population. When items are modified, however, they no longer retain the same psychometric properties. Scale alignment is necessary to make the inferences from test scores comparable across different forms. However, the reality is that the sample size of the students taking modified test forms is typically so small that it precludes direct calibration or linking procedures. This calls for alternative approaches to ensure test form comparability.

One intuitive and widely used approach stems from the following observation. Certain items or task types are straightforward to modify into an acceptable format, depending on the nature of the disability. They require less extensive changes than other task types. Naturally, if one chooses to only modify those items or task types that have readily available modifications, and assume the original item parameters calibrated in the general student population remains unchanged, one could avoid being asked the question of comparability, at least on the surface.

We find this approach somewhat problematic from a test validity and fairness point of view. It shrinks the pool of item or task types available for modification, which leads to a reduction in standards, attributes, or content that can potentially appear in the modified test forms. Leaving those items or tasks out results in a potentially fundamentally different test because students with disabilities are now being presented fewer opportunities to demonstrate their learning and achievement than students who can take the unmodified version of the test. Therefore, we need a more open-minded approach that permits the modification of as many items or task types as possible, but it should be accompanied by a rigorous methodology to evaluate the impact of the changes on the items to ensure comparability of the scores.

As a response, we develop a systematic method that integrates the structured judgments provided by subject matter experts with known psychometric properties of the unmodified test to evaluate the impact of the modifications and to provide adjustments to the items parameters. The utilization of the subject matter experts' input is a key component of our method, which has a cross-classified random effects model at its core. These expert judges are the ones who work most

closely with the students who require additional support. They practice adaptation and modification on the day-to-day basis to aid instruction and learning, and yet their voices have so far been largely left out in the context of summative educational assessment development. We felt that the field has an obligation to conduct research that broadens perspectives, empowers the learning of all students, helps the most vulnerable students overcome barriers and disabilities, and provides opportunities so that all students can demonstrate their achievement.

1.1. Motivating Example

One particular area in which these equity and comparability issues are most pronounced is the English language proficiency (ELP) assessment for students who are blind or have low vision (BLV). The English Language Proficiency Assessment for the 21st Century (ELPA21), a large multistate ELP assessment program, developed a Braille version of their main online K–12 ELP assessment for BLV students. Specifically, from a pool of online items for ELs from the general population, a subset of items were chosen and modified into a Braille version. Students in the BLV population take modified Braille test forms. For item analysis and test scoring, ELPA21 uses item response theory (IRT). The item parameters calibrated with responses to the online assessment from the general population are inappropriate for the Braille test items unless we make proper adjustments to the original item parameters because of the sometimes extensive changes that must be made (Winter et al., 2018).

Unfortunately, neither standard linking methods nor IRT item calibration is possible in the present situation. The very small sample size hinders item calibration. In the 2015-16 ELPA21 summative assessment, for example, a total of 26 students, which amounts to only .007% of all tested students in ELPA21, took modified Braille test forms. Discussions of the sample size issue in the IRT literature suggest that only with a large sample size can the item parameters be accurately estimated, which, in turn, improves scaled score estimation (DeMars, 2010). Although there is no gold standard (Morizot et al., 2007), and the sample size depends heavily on the nature of the IRT model, there appears to be an agreement that a sample size less than 100 is unacceptable even for the one-parameter logistic model (Embretson & Reise, 2000). Given that the number of BLV students is considerably smaller than 100, IRT item calibration is not an option.

If IRT calibration is not feasible, one may wonder about the appropriateness of using a standard linking procedure. In the linking literature, two approaches are frequently used: the common population approach and the common item approach (Yamamoto & Mazzeo, 2005). Obviously, common population linking is not appropriate because there does not exist a group of ELs that can take both the original online test and the BLV test. Common item linking is also

inappropriate because when extensive modifications are involved, there is no assurance that the items in common (the "anchors") have similar item parameters. Even if the item parameters of the modified items in the Braille version and their source in the original online version were somehow similar, the small sample size again hinders any attempt at statistical confirmation.

1.2. A New Linking Methodology

To address the challenge, we utilize data generated by a new linking methodology, namely a "judgement-based, data-informed linking process" developed by Winter et al. (2018) for linking the two seemingly unlinkable tests. For the sake of simplicity, we limit our discussion to dichotomously scored items. In Winter et al. (2018), convened panel of experts provided the expected probabilities of correct responses to modified items from special population at two cut scores. The two cut scores are anchored by detailed proficiency-level descriptors and represent the two most important decision points along the proficiency scale that contribute to overall proficiency classifications. The judges were informed of the extent of the modification. Their judgments of the modified items were also anchored to the actual correct response probabilities of the original items at the cut scores. The last point is unique to this approach.

In terms of determining item statistics, Winter et al.'s (2018) method did not come entirely out of the blue. It evokes resemblance to anchor-based methods for judgmentally estimating item statistics (Hambleton & Jirka, 2006). The literature indicates that the judgment-based approach is generally a promising one (e.g., Farmer, 1928; Lorge & Diamond, 1954; Lorge & Kruglov, 1952; Thorndike, 1982). Even though Hambleton and Jirka (2006) suggested the usefulness of involving judges to estimate item difficulty, Winter et al. (2018) may be the first to directly utilize the judgment-based approach for the purpose of linking. In turn, from the linking perspective, Winter et al.'s (2018) method is related to scale alignment by "social moderation" (Linn, 1993; Mislevy, 1993).

On the other hand, Winter et al.'s (2018) method is similar to standard setting procedures. Indeed, standard setting and linking are closely related. Newly developed test forms often lead to differences in difficulty; hence, appropriate linking is needed to use the same cut scores. When test linking is not feasible, they may be replaced with standard resetting to maintain the "performance standard" (Bramley & Benton, 2017; Dwyer, 2016). Specifically, Winter et al.'s (2018) method resembles the Angoff standard setting method but with a different purpose and a unique feature. The Angoff (1971) method involves expert panelists who review each item and provide estimated proportions of correct responses for a population of interest (see also Cizek & Bunch, 2007). While the Angoff method is a standard setting procedure to determine cut scores on the proficiency scale, the purpose of Winter et al.'s (2018) method is the adjustment of item statistics for the modified test forms so that comparable scores may be obtained.

The unique feature is the availability of the proportions of correct responses from the online assessment administered to the general population. Thus, this new method may be deemed an "anchored" Angoff method.

If one tries to set new cut scores, however, the modified test score scale cannot be interpreted in the same manner as the original. By providing estimates of item parameters directly, we enhance the comparability of scores without having to set new cut scores. With new item parameters, we are able to provide scores and profiles of proficiency on an equal footing across all population with the equivalent cut scores used for the original test forms.

We augment Winter et al.'s (2018) method with a comprehensive approach to estimate new item parameters. Existing methods utilizing raters' judgments have relied on fixed-effects linear models (e.g., averages of the ratings) or pursued consensus among raters through multiple rounds of judgment. Instead, we propose *moderated item calibration* through the application of a cross-classified random effects model that simultaneously utilizes the panelists' ratings and the data from the general population item calibration. It is called "moderated" item calibration because the item parameters for modified items are estimated from two sources, once from expert judgments like in "social moderation" and twice from student data from the general population.

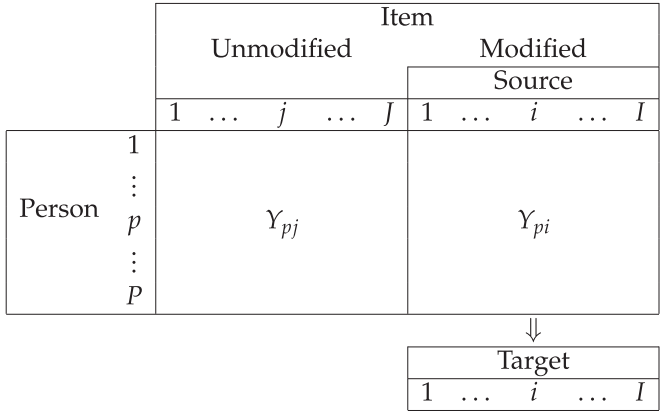
Before introducing our approach, we first note that Bolsinova et al. (2017) has also developed a method for test linking utilizing expert judgments. In their study, expert judgments are incorporated as priors in Bayesian estimation to combine with the linking data, which is sparse. Our approach proposed herein based on this new linking methodology is different from their approach. First, as described in Subsection 1.1, we assume a situation where the linking data set is not available. Bolsinova et al. (2017), on the other hand, considers a pretest nonequivalent group equating design with linking groups, a scenario where linking data are available though it may not be large enough to guarantee high quality of linking. Second, the manner in which we utilize the expert judgments differs. In our study, the expert judgements directly provide the basis for the item difficulty of the new test forms. On the other hand in Bolsinova et al.'s (2017) study, the expert judgments were used as informative priors for capturing the difference in the (average) difficulty of items.

2. Cross-Classified Random Effects Model for Moderated Item Calibration

2.1. Data Structure

We consider the data as consisting of two parts: the original calibration data (from the general population's responses to the original test items) and the judge-provided data for modified test items (from a panel of expert raters). Recall that we only consider dichotomous responses, with no loss of generality. As a departure from the fixed-item tradition in item response modeling, we regard the design of the original calibration data as cross-classified with students/persons

TABLE 1.
Original Calibration Data Structure



crossed with items. Suppose there are $p = 1, ..., P$ independent persons, $j = 1, \ldots, J$ *unmodified* items, and $i = 1, \ldots, I$ *modified* items. *Modified* items are distinguished between *source* items and *target* items. Relevant to the original calibration data are source items. *Source* items refer to those items in the original online assessment that have been chosen for modification. In other words, they are the "source" or "parent" items (from the original assessment) that are subsequently modified for the special population. *Unmodified* items are the remaining online-only items.

Let Y_{pj} and Y_{pi} be Bernoulli random variables, indicating the response from person p to unmodified item j and source item i, respectively. Table 1 presents the structure of the item response data. For convenience, we group "unmodified" items into one block and "source" items into another. Each block contains a crossed person-by-item design. The $P \times (J + I)$ matrix shown in the top part of Table 1 represent the responses to the online items from the general population. The block on the bottom contains *target* items, which refer those test items to be presented to special population, after modification. There is a one-to-one correspondence between those 'source' items and the "target" items, as indicated by the use of the shared index variable $i = 1, \ldots, I$.

The second part of the data involves the judge-provided ratings on the "target" items. The raters ($r = 1, \ldots, R$) provide $k = 1, \ldots, K$ ratings, at the K proficiency level cut-offs, for each target test item i. Let $Y_{rik} \in [0, 1]$ denote the kth rating from rater r to item i. It is important to note that the raters provide a judgment on the 0% to 100% probability scale. They are informed of the proportion of correct responses to the source items, based on item calibration in the general population. Table 2 shows the structure of the judge-provided data. It is a

TABLE 2.
Judge Data Structure

		Target				
		1	...	i	...	I
Rater	1					
	⋮					
	r			Y_{ri1} ⋮ Y_{riK}		
	⋮					
	R					

two-way table with raters as rows and items as columns, that is, a crossed $r \times i$ design, containing K repeated measures in each cell.

2.2. Model Formulation

We now construct a three-part model: two for the original calibration data and one for the judge-provided data, with the goal of estimating revised item parameters for the target test items through *moderated item calibration*.

2.2.1. Part I. The first model is a familiar one. For the unmodified items of the original calibration data, represented by the $P \times J$ block on the left side of Table 1, we use a two-parameter logistic (2PL) model, following ELPA21's choice in operational item calibration. Let θ_p be the latent proficiency variable for person p. In the context of ELPA21, this could be the proficiency in one of the four language domains: listening, speaking, reading, and writing. Conditional on θ_p, the response probability of person p to item j is

$$\text{logit}\left[P(Y_{pj} = 1|\theta_p)\right] = a_j^*\theta_p + c_j^*, \tag{1}$$

where a_j^* and c_j^* are, respectively, the item slope parameter and the intercept for unmodified item j. The operational item parameters are assumed known, and we use asterisks to indicate that they are fixed values. Because the unmodified items do not serve as source items for any BLV modification, updating their item parameters is not of interest. Again, adopting ELPA21's operational assumption of the general population proficiency distribution, we assume $\theta_p \sim N(\mu_\theta, \sigma_\theta^2)$. When the original calibration were conducted, the population mean and variance were fixed to 0 and 1, respectively, for identification. When fixed item parameters are applied to data other than the original calibration sample, one can

estimate μ_θ and σ_θ^2 as free parameters. In Part I, $\boldsymbol{\omega}_1 = (\mu_\theta, \sigma_\theta^2)'$ contains the structural parameters.

In essence, our Part I model resembles IRT-based fixed item parameter equating. The item responses, in conjunction with the operational item parameters, place the individuals' latent proficiency variables on a known scale. Instead of computing estimates of θ_p, we directly incorporate them as random variables in the second part of the model.

2.2.2. Part II. The second model is for the source items chosen for modification, represented as the $P \times I$ block on the right side of Table 1. To further simplify subsequent analysis and reporting, we assume that the item parameters of the (yet to be determined) modified items or target items are connected to the source items, through adjustments to the item difficulty parameters. The judges inform the adjustments via their ratings.

The model we use is a cross-classified random person and random item model. Instead of making the item parameters fixed, as in standard IRT calibration, we make the item intercepts random effects. This facilitates subsequent development. The response probability of person p to item i is modeled as

$$\text{logit}\left[P(Y_{pi} = 1|\theta_p, c_i)\right] = a_i^*\theta_p + c_i, \tag{2}$$

where a_i^* is the item slope parameter and $c_i \sim N(\mu_c, \sigma_c^2)$ is the random effects intercept term for modified item i.

The asterisk on a punctuates the fact that a_i^* is not estimated but rather fixed to the original calibration slope, per the aforementioned assumption that adjustments derived from the ratings affect difficulty. Importantly, θ_p is directly imported from Part I because the students are the same, from the general population. In Part II, $\boldsymbol{\omega}_2 = (\mu_c, \sigma_c^2)'$ contains the structural parameters.

2.2.3. Part III. We now connect raters' judgment with information from the original general population calibration. The variability in raters' judgment can be decomposed into two general components: the variability in item characteristics (difficulty in this case) and the variability between and within raters. The model has cross-classified random effects (raters and items), is nonlinear (due to the sigmoidal shape of item response functions in IRT), and is multivariate (due to the presence of judgments at multiple cut scores). Following Raudenbush and Bryk's (2002) notation, we present the model as a two-level hierarchical model with a "within-cell" model and a "between-cell" model. As shown in Table 2, the cells are formed by crossing raters with items.

Within-cell model: In each cell, we have repeated measures of ratings. Recall that $Y_{rik} \in [0, 1]$ is the kth rating in probability scale from rater r on item i. At cut score $k = 1, \ldots, K$, denoted θ_k^*, the within-cell model can be written as

$$Y_{rik} = \frac{1}{1 + \exp\left[-(a_i^*\theta_k^* + \eta_{ri})\right]} + e_{rik}, \tag{3}$$

where η_{ri} is interpretable as the mean (in logit) in cell *ri*, and e_{rik} is the random error and/or interaction effect between rater and item. Collecting terms, we assume the vector $\mathbf{e}_{ri} = (e_{ri1}, \ldots, e_{riK})'$ follows a multivariate normal distribution $\mathbf{e}_{ri} \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_e)$, where $\boldsymbol{\Sigma}_e$ is a general covariance matrix. It is important to note that while the model resembles an IRT model, the judges were asked to provide estimates of correct response probabilities at fixed cut score levels θ_k^*. With the fixed item slopes, the K regression lines in Equation 3 differ by known *offset* values $(a_i^*\theta_k^*)$, adopting standard terminology from generalized linear models (McCullagh & Nelder, 1989).

Between-cell model: Next, we model variations among the ratings (between cells) by splitting them into item and rater components. Again following the tradition in hierarchical linear models, we present the "between-cell" model in two separate equations: (1) the unconditional "between-cell" model for the linear predictor η_{ri} and (2) the item regression equation or the so-called means-as-outcomes regression equation. The unconditional between-cell model can then be written as

$$\eta_{ri} = \alpha_i + \gamma_r, \tag{4}$$

where α_i is the item random effect and γ_r is the rater random effect. This resembles a classical two-way crossed analysis of variance model. The raters are the rows and the items are the columns in the two-way layout. We assume the rater random effects follow a normal distribution $\gamma_r \sim N(0, \sigma_\gamma^2)$. As to the items, we characterize the item random effect with the following means-as-outcomes regression equation that adjusts the target items' location parameters with random item intercepts from Part II:

$$\alpha_i = \beta_0 + \beta_1 c_i + \zeta_i, \tag{5}$$

where β_0 is the fixed intercept, β_1 is the fixed regression coefficient, and the error term is normally distributed $\zeta_i \sim N(0, \sigma_\zeta^2)$. In Part III, $\boldsymbol{\omega}_3 = (\text{vech}(\boldsymbol{\Sigma}_e), \beta_0, \beta_1, \sigma_\zeta^2, \sigma_\gamma^2)$ contains the structural parameters.

Equation 5 is a latent variable regression because both α_i and c_i are random effects/latent variables. This regression equation may be viewed as a latent variable counterpart to the item parameter based equating approaches discussed in the earlier test equating literature (e.g., Bejar & Wingersky, 1982). The β_0 and β_1 values are the "linking constants" that attempt to put the parameters from the general population item calibration on the scale of the special population. We will discuss how estimates of β_0 and β_1, along with other structural parameters, can be obtained from data. Revised item difficulty parameters for the target items could be based on empirical Bayes (EB) predictions of the random effect α_i.

Combined model: Substituting elements from Equations 4 and 5 into Equation 3, we obtain a multivariate nonlinear cross-classified random effects model:

$$Y_{rik} = \frac{1}{1 + \exp\left[-(a_i^*\theta_k^* + \beta_0 + \beta_1 c_i + \zeta_i + \gamma_r)\right]} + e_{rik}. \tag{6}$$

We term this approach *moderated item calibration.* The target items' parameters are calibrated with information contributed by the calibration data, albeit indirectly. These item parameters are also determined by ratings from expert panels, but then the expert judgments are moderated by student data from the general population. Again, one can see the item parameters for target items are estimated from two sources of moderation, one from expert judgments like in 'social moderation' and another from student data from the general population.

2.3. Conditional Distributions and Likelihoods

2.3.1. Part I. The conditional distribution of Y_{pj} is Bernoulli:

$$f(y_{pj}|\theta_p) = \left[P(Y_{pj} = 1|\theta_p)\right]^{y_{pj}}\left[1 - P(Y_{pj} = 1|\theta_p)\right]^{1-y_{pj}}. \tag{7}$$

Let $\mathbf{Y}_1$ be the $P \times J$ matrix of item responses to all J unmodified items from all P persons (the first block in Table 1). Next, we need to invoke the conditional independence assumption for the responses to the unmodified items (conditional on θ_p), as well as the independence of the persons. We see that

$$f(\mathbf{Y}_1|\boldsymbol{\theta}) = \prod_{p=1}^{P}\prod_{j=1}^{J} f(y_{pj}|\theta_p), \tag{8}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)'$ contains all the latent proficiency variables.

2.3.2. Part II. The conditional distribution of Y_{pi} is again Bernoulli:

$$f(y_{pi}|\theta_p, c_i) = \left[P(y_{pi} = 1|\theta_p, c_i)\right]^{y_{pi}}\left[1 - P(y_{pi} = 1|\theta_p, c_i)\right]^{1-y_{pi}}. \tag{9}$$

Let $\mathbf{Y}_2$ be the $P \times I$ matrix of item responses to all I source items from all P persons. Once again, given θ_p and c_i, we assume the conditional independence of item responses to the source items chosen for modification, which leads to

$$f(\mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c}) = \prod_{p=1}^{P}\prod_{i=1}^{I} f(y_{pi}|\theta_p, c_i), \tag{10}$$

where $\mathbf{c} = (c_1, \ldots, c_I)'$ contains the random item intercepts.

Combining Equations 8 and 10, and making the assumption of the conditional independence of responses to the J unmodified items and I source items on θ and $\mathbf{c}$, we see that

$$f(\mathbf{Y}_1, \mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c}) = f(\mathbf{Y}_1|\boldsymbol{\theta})f(\mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c}). \tag{11}$$

Note that the above conditional independence assumptions are always made in ELPA21 operational calibration studies that fit 2PL IRT models to the $I + J$ items jointly. ELPA21 does not utilize the random intercept formulation in our Part II model. Instead, fixed intercept parameters are specified for each item, along with fixed slope parameters, so that an item bank can be maintained using standard methods.

2.3.3. Part III. Let $\mathbf{y}_{ri} = (Y_{ri1}, \ldots, Y_{riK})'$ denote a $K \times 1$ random vector of ratings in a cell. In addition, let $\mathbf{e}_{ri} = (e_{ri1}, \ldots, e_{riK})'$ be the corresponding $K \times 1$ random vector of the error terms. The combined model in Equation 6 can be reexpressed in matrix form as $\mathbf{y}_{ri} = g(\boldsymbol{\phi}_{ri}) + \mathbf{e}_{ri}$, where the kth element of $g(\boldsymbol{\phi}_{ri})$ is defined as $g(\phi_{rik}) = 1/(1 + \exp[-\phi_{rik}])$, and

$$\phi_{rik} = a_i^* \theta_k^* + \beta_0 + \beta_1 c_i + \zeta_i + \gamma_r.$$

The conditional distribution of $\mathbf{y}_{ri}$ given c_i, ζ_i, and γ_r comes from the combined model in Equation 6:

$$f(\mathbf{y}_{ri}|c_i, \boldsymbol{\zeta}_i, \boldsymbol{\gamma}_r) = |2\boldsymbol{\pi}\boldsymbol{\Sigma}_e|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[\mathbf{y}_{ri} - g(\boldsymbol{\phi}_{ri})]'\boldsymbol{\Sigma}_e^{-1}[\mathbf{y}_{ri} - g(\boldsymbol{\phi}_{ri})]\right\}. \quad (12)$$

Let $\mathbf{Y}_3$ denote all the judge ratings in the $R \times I$ cells. We assume that conditional on c_i, ζ_i, and γ_r, the cell probabilities factor:

$$f(\mathbf{Y}_3|\mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = \prod_{i=1}^{I}\prod_{r=1}^{R} f(\mathbf{y}_{ri}|c_i, \zeta_i, \gamma_r), \quad (13)$$

where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_I)'$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_R)'$.

2.4. The Likelihood

Equations 11 and 13 are connected by their shared dependence on the item random effects $\mathbf{c}$. Assuming conditional independence on $\mathbf{c}$, we can write down a joint conditional distribution of the entire observed item response data (used for original calibration) and the entire set of rater judgments:

$$f(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = f(\mathbf{Y}_1, \mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c}) f(\mathbf{Y}_3|\mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma}), \quad (14)$$

where $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3)$ indicates all the observed data. Equation 14 reflects a latent variable model that relies on $f(\mathbf{Y}_1, \mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c})$ (Part I and Part II models) to measure the latent variables $\mathbf{c}$ (source item random intercepts) and then introduce $\mathbf{c}$ as a latent predictor in $f(\mathbf{Y}_3|\mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma})$ (Part III model), with the item and rater residual effects represented by $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$ in a two-way crossed random effect model for the ratings.

Treating $\boldsymbol{\theta}$, $\mathbf{c}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\gamma}$ as missing data, we see that the complete data likelihood function is

$$L(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = f(\mathbf{Y}_1, \mathbf{Y}_2|\boldsymbol{\theta}, \mathbf{c}) f_{\omega_3}(\mathbf{Y}_3|\mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) f_{\omega_1}(\boldsymbol{\theta}) f_{\omega_2}(\mathbf{c}) f_{\omega_3}(\boldsymbol{\zeta}) f_{\omega_3}(\boldsymbol{\gamma}), \quad (15)$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3)$ contains all the structural parameters, and the prior distributions on $\boldsymbol{\theta}$, $\mathbf{c}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\gamma}$ all take factored forms, with subscripts denoting appropriate dependence on $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$. This complete data likelihood will be utilized to derive maximum marginal likelihood estimates of $\boldsymbol{\omega}$.

3. Estimation of Structural Parameters and Prediction of Random Effects

3.1. Parameter Estimation via the Metropolis–Hastings Robbins–Monro (MH-RM) Algorithm

MH-RM is a data-augmented RM algorithm (Robbins & Monro, 1951) coupled with the MH algorithm (Hastings, 1970; Metropolis et al., 1953). This hybrid algorithm was first proposed by Cai (2008, 2010a, 2010b). It is based on two insights: Fisher's (1925) identity and the second is rooted in Robbins and Monro's (1951) classical stochastic approximation method. The method has since been used successfully in a number of different modeling contexts (Cai, 2015; Falk & Cai, 2016; Monroe, 2014; Monroe & Cai, 2014; Yang & Cai, 2014).

Let $\boldsymbol{\omega}^{(t)}$ be the parameter estimates at iteration t. At iteration $t+1$, the MH-RM algorithm follows three steps: *stochastic imputation, stochastic approximation, and the RM update*.

Step 1: Stochastic imputation

Draw m_t sets of missing data $\{\mathbf{M}_j^{(t+1)}; j = 1, \ldots, m_t\}$ using the MH sampler with the posterior predictive distribution $\Pi(\mathbf{M}|\mathbf{Y}, \boldsymbol{\omega}^{(t)})$ of missing data. Then m_t sets of complete data $\{\mathbf{Y}, \mathbf{M}_j^{(t+1)}; j = 1, \ldots, m_t\}$ are created. In our context, $\mathbf{M} = (\boldsymbol{\theta}, \mathbf{c}, \boldsymbol{\zeta}, \boldsymbol{\gamma})$.

Step 2: Stochastic approximation

Based on the imputed data, the ascent directions are determined by evaluating the complete data log-likelihood and its gradients. The gradient of the complete data log-likelihood is

$$\mathbf{s}(\boldsymbol{\omega}^{(t)}|\mathbf{Y}, \mathbf{M}_j^{(t+1)}) = \frac{\partial}{\partial \boldsymbol{\omega}} l(\boldsymbol{\omega}^{(t)}|\mathbf{Y}, \mathbf{M}_j^{(t+1)}). \tag{16}$$

In practice, we can approximate it with the sample average of the complete data gradients:

$$\tilde{\mathbf{s}}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} \mathbf{s}(\boldsymbol{\omega}^{(t)}|\mathbf{Y}, \mathbf{M}_j^{(t+1)}). \tag{17}$$

By virtue of Fisher's (1925) identity, the conditional expectation of the gradient of the complete data log-likelihood over the posterior distribution of missing data is equal to the gradient of the observed data (marginal) log-likelihood:

$$\frac{\partial}{\partial \boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y}) = \int \mathbf{s}(\boldsymbol{\omega}|\mathbf{Y}, \mathbf{M}) \Pi(\mathbf{M}|\mathbf{Y}, \boldsymbol{\omega}) d\mathbf{M}. \tag{18}$$

This guarantees that the ascent direction is correct on average. In addition, we compute the conditional expectation of the complete data information matrix, the purpose of which is to improve stability and speed. For complete data information matrix

$$\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y}, \mathbf{M}) = -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y}, \mathbf{M})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'}, \tag{19}$$

we can also compute its Monte Carlo approximation as

$$\mathbf{H}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} \mathbf{H}(\boldsymbol{\omega}^{(t)}|\mathbf{Y}, \mathbf{M}_j^{(t+1)}). \tag{20}$$

Step 3: The RM update

The RM filter is applied as we update the parameter estimates. It is

$$\boldsymbol{\omega}^{(t+1)} = \boldsymbol{\omega}^{(t)} + \epsilon_t (\Gamma_{t+1}^{-1}) \tilde{\mathbf{s}}_{t+1}, \tag{21}$$

where Γ_{t+1}^{-1} is a recursive stochastic approximation of the conditional expectation of the complete data information matrix. Given initial values $(\boldsymbol{\omega}_0, \Gamma_0)$, where Γ_0 is a symmetric positive definite matrix, a recursive approximation of E $(\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y}, \mathbf{M}))$ is defined as

$$\Gamma_{t+1} = \Gamma_t + \epsilon_t (\mathbf{H}_{t+1} - \Gamma_t), \tag{22}$$

where ϵ_t is a sequence of gain constants such that $\epsilon_t \in (0, 1]$, subject to $\sum_{t=0}^{\infty} \epsilon_t = \infty$ and $\sum_{t=0}^{\infty} \epsilon_t^2 < \infty$. Essentially, the role of ϵ_t is to eliminate the noise effect introduced by imputing for missing data.

In practice, the iterations are divided into multiple stages to improve stability and speed. In Stage I, we run some iterations to bring the starting values to the general neighborhood of the parameter estimates. Further iterations are subsequently performed in Stage II, wherein the averages of the estimates in the neighborhood become the starting values for in Stage III. Accordingly, multistage gain constants are favored for ϵ_t such that gain constants with fixed values are set in Stages I and II, followed by decreasing gain constants in Stage III (Cai, 2008, 2015). Note that the gain constants can be determined after monitoring the traces of parameters. Iterations over the three steps are terminated when the minimum successive differences of the estimates for a predetermined window size reaches a convergence criteria.

Recall that we need to draw plausible values of $\mathbf{M}_j^{(t+1)}$ from its posterior predictive distribution of missing data $\Pi(\mathbf{M}|\mathbf{Y}, \boldsymbol{\omega}^{(t)})$ in Step 1 of the MH-RM workflow. To impute $\mathbf{M}$, a Metropolis-within-Gibbs algorithm by Patz and Junker (1999) shall be used, following Cai (2008, 2010a, 2010b). The sampling of $\boldsymbol{\theta}$,

$\mathbf{c}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\gamma}$ are done in alternation. A key insight is that the imputation of each set of missing data is made easier by conditioning on all the other sets. The strategy is quite similar to the imputation posterior or alternating imputation posterior algorithm (Cho & Rabe-Hesketh, 2011; Clayton & Rasbash, 2011).

3.2. Estimation of Standard Errors (SEs)

*SE*s are also available under MH-RM. Two types of *SE*s, named after the methods used to obtain them, have been proposed: (1) recursively approximated *SE*s and (2) postconvergence approximated *SE*s (Yang & Cai, 2014). *SE*s are generally obtained by the square root of the diagonal elements of the inverse of the observed data information matrix. Both approaches follow Louis (1982) for approximating the observed data information matrix. They have been used and examined in a number of research studies for various models (e.g., Cai, 2010a; Falk & Cai, 2016; Monroe & Cai, 2014; Yang & Cai, 2014). Herein, we present the postconvergence approach, as it is studied in our simulation.

According to Louis (1982), the information matrix of the observed data log-likelihood is

$$-\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} = \mathrm{E}(\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})) - \mathrm{E}(\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})[\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})]') + \mathrm{E}(\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M}))\mathrm{E}([\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M}]'). \tag{23}$$

This approach directly applies Louis's (1982) equation, wherein Monte Carlo integration is used upon convergence of MH-RM (Diebolt & Ip, 1996). The convergence with respect to the maximum likelihood (ML) estimate $\hat{\boldsymbol{\omega}}$ implies $\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M}) = \mathbf{0}$, and accordingly the last term in Equation 23 no longer exists when evaluated at $\hat{\boldsymbol{\omega}}$. Let m_c denote the number of samples drawn to approximate the information matrix. The first term in Equation 23 is approximated as

$$\mathrm{E}(\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})) \approx \frac{1}{m_c}\sum_{j=1}^{m_c}\mathbf{H}(\hat{\boldsymbol{\omega}}|\mathbf{Y},\mathbf{M}_j), \tag{24}$$

and similarly, the second term is computed as

$$\mathrm{E}(\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})[\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{M})]') \approx \frac{1}{m_c}\sum_{j=1}^{m_c}(\mathbf{s}(\hat{\boldsymbol{\omega}}|\mathbf{Y},\mathbf{M}_j)[\mathbf{s}(\hat{\boldsymbol{\omega}}|\mathbf{Y},\mathbf{M}_j)]'). \tag{25}$$

3.3. Prediction of Random Effects

We use EB prediction given maximum marginal likelihood estimate $\hat{\boldsymbol{\omega}}$ of the structural parameters (McCulloch et al., 2008; Skrondal & Rabe-Hesketh, 2014). Specifically, an EB estimate of the random effects $\mathbf{M}$ is defined as $\mathrm{E}(\mathbf{M}|\mathbf{Y};\hat{\boldsymbol{\omega}})$, which is derived from the following conditional posterior predictive distribution.

$$\Pi(\mathbf{M}|\mathbf{Y};\hat{\boldsymbol{\omega}}) = \frac{f(\mathbf{Y}|\mathbf{M};\hat{\boldsymbol{\omega}})\Pi(\mathbf{M})}{\int f(\mathbf{Y}|\mathbf{M};\hat{\boldsymbol{\omega}})\Pi(\mathbf{M})d\mathbf{M}}, \tag{26}$$

While our main interest is in the determination of the item random effects c_i and ζ_i to compute α_i so that corrected item parameters may be obtained for the special population, we present scoring for all random effects with generic notation. Given $\hat{\omega}$, we draw additional samples after convergence to compute sample averages of $\mathbf{M}$:

$$\hat{\mathbf{M}} \approx \frac{1}{m_c}\sum_{j=1}^{m_c}\mathbf{M}_j. \tag{27}$$

This can be done in conjunction with the postconvergence *SE* estimation.

4. Empirical Application

We apply the proposed model and estimation methods to data from ELPA21. The results from this empirical application will also inform our simulation studies.

4.1. Data and Methods

The calibration data set of the original item pool is from the 2015–2016 field test. The ELPA21 summative assessment has four domains (listening, reading, writing, and speaking) and consists of six gradebands (Grade K, Grade 1, Grade 2–3, Grade 4–5, Grade 6–8, and Grade 9–12). There are multiple test forms per gradeband that were randomly assigned to students with a planned missing data design to maximizing the number of test items that can be field tested. The missing data are missing completely at random (Rubin, 1987). Sample size ranges from approximately 37,000 to 55,000. The number of online items per domain ranges from approximately 40 to approximately 160.

The judge-provided rating data were collected in a cut score linking panel meeting held in 2018. From data generated by the panel meeting, we selected the Listening domain for our application because most modified Braille items for listening are dichotomous. For each domain and gradeband, there are typically dozens of BLV items. There are three to five raters examining these BLV items, but in most cases, there are four raters. Multiple rounds of judging data were collected. For our analysis, we randomly selected 10,000 students from the general population on the original test items, and we only used the judge ratings from the first round. As a reminder, raters provided ratings at two cut scores so $K = 2$ per Item $\times$ Rater combination.

We used an MH-RM algorithm coded and implemented in R by the authors (R Core Team, 2018) to jointly estimate the unified model (Part I + Part II + Part

III). The number of Stage I iterations was set to 3,000, and 500 Stage II iterations were used with fixed gain constants equal to 0.1. At Stage III, the decreasing gain constants $\epsilon_t = \epsilon_0 / t^{0.75}$ were applied, where ϵ_0 are the gain constants used in Stages I and II. A "burn-in" of 10 was used for an MH sampler, and simulation size m_t was set to 1 across the stages. Convergence was determined by a window of three iterations within a 1.0×10^{-4} tolerance threshold. After convergence, $25,000$ Monte Carlo draws were used to score all random effects/factor. Post-convergence *SE*s were computed from 250 draws with a thinning interval of 20.

In addition, we compared our approach against a linear fixed-effects approach. For dichotomous data, we have two simple regression equations whose outcome variable is the mean expected (logit) probability of correct responses to the modified Braille test items judged by raters. Note that in this approach, the item parameters are determined by the expert ratings with no moderation by item parameters in the general population.

4.2. Results

The results of the parameter estimates and *SE*s are presented in Table 3. First, the original calibration data were used, so we simply set the mean and variance of θ_p in the Part I model to 0 and 1, respectively. Thus there is no free parameter from Part I to report. From the positive means of the random item intercepts for the original source items in Part II, we find that the items are relatively easy overall. There is also appreciable variability in the item's difficulty.

From the Part III results, we see negative intercepts and regression coefficient largely close to 1. Recall that the intercept β_0 and the regression coefficient β_1 are the "linking constants." Specifically, β_0 is interpreted as the grand mean item difficulty (location) over the items and raters, and β_1 is the effect of (random) item difficulty of the original source items. With $\beta_1 = 1$, interpretation of both parameters is straightforward. The negative intercept suggests that the BLV items tend to be more difficult than the original source items, which of course has been determined by the judges. In other words, there are constant downward shifts in item location for all items, with the variations among the items essentially remaining the same as in the original source items. With β_1 not close to 1, the difference between the item difficulties of the original source items and modified Braille items varies much more by item. Take an example of four items with the difficulties of the original items being 2.5, 1.5, $-.5$, and -1; and β_0 and β_1 .5 and .6, respectively. The resulting revised item difficulty estimates become 2.0, 1.4, $-.1$, and .2. As seen here, we cannot assume equivalent effect on changes in item difficulty. This is hardly surprising given what is demanded of the BLV students while taking a test of ELP in Braille, in comparison to the general population. This study, however, is the first to present a rigorous empirical means to quantify the magnitude of the difference.

TABLE 3.
Parameter Estimates and Standard Errors

		Part II		Part III						
Domain	Grade	μ_c	σ_c^2	β_0	β_1	σ_ξ^2	σ_γ^2	$\Sigma_e^2(1,1)$	$\Sigma_e^2(2,1)$	$\Sigma_e^2(2,2)$
Listening	K	0.73816	0.53345	−0.23592	1.04232	0.01240	0.06810	0.0024	0.00174	0.00223
		(.16729)	(.17315)	(.04689)	(.05028)	(.00398)	(.05068)	(.00045)	(.00038)	(.00041)
	1	2.02460	0.94651	−0.07494	1.01878	0.00014	0.00444	0.00173	0.00200	0.00259
		(.17942)	(.25857)	(.00012)	(.00005)	(.00000)	(.00358)	(.00026)	(.00030)	(.00036)
	2–3	1.81866	0.97784	−0.68437	1.10077	0.06373	0.03362	0.00601	0.00435	0.00661
		(.20093)	(.28244)	(.25804)	(.09712)	(.01674)	(.03315)	(.00118)	(.00111)	(.00129)
	4–5	1.68685	1.47547	−0.50493	0.98319	0.03064	0.00013	0.00506	0.00441	0.00534
		(.25504)	(.43726)	(.12496)	(.05484)	(.01130)	(.00014)	(.00098)	(.00106)	(.00096)
	6–8	1.66366	1.18355	−0.32421	1.06199	0.03032	0.03559	0.00411	0.00279	0.00397
		(.20507)	(.31775)	(.05677)	(.04338)	(.00685)	(.02471)	(.00056)	(.00054)	(.00055)
	9–12	1.31879	1.36448	−0.18640	1.03107	0.00108	0.00390	0.00166	0.00224	0.00352
		(.21343)	(.35209)	(.00135)	(.00078)	(.00002)	(.00283)	(.00023)	(.00030)	(.00042)

Note. Standard errors are in parentheses. We present the values up to five decimal points due to the small magnitudes of the variance estimates, which result from small units of the probability scale.

TABLE 4.
Comparison With Linear Fixed-Effects Approach

	Mean	*SD*	Minimum	Maximum
K	.078	.211	−.272	0.730
1	.235	.417	−.246	1.799
2–3	.128	.356	−.353	1.027
4–5	.318	.512	−.068	2.365
6–8	.265	.320	−.016	1.351
9–12	−.006	.102	−.202	0.256

Also of interest is the variance of residual random item effects and rater effects. The small variance estimates overall are due to the constraint of the probability scale. But by including the original item location parameters, a substantial amount of variance in the judge-determined item locations is explained away. There appears to be a strong relationship between the judge-determined difficulty and the original source item's difficulty estimated from item response data. This finding is also encouraging in that it suggests linking efforts such as this may yield fruitful results.

We compared the item location parameters obtained from our approach against those from the simple linear fixed-effects approach. Table 4 displays the differences in the item location parameters from the two approaches in terms of mean, standard deviation (*SD*), minimum, and maximum. The magnitude of difference varies by gradeband and also by items within each gradeband. The results indicate that applying our approach makes a meaningful difference for several items compared to using the simpler approach.

5. Simulation Study

5.1. Simulation Study Design

We selected simulation conditions such that they both reflect and complement the real conditions in ELPA21. Since the primary interest lies in inclusion of the judge-provided data into otherwise standard IRT models, we manipulated conditions for the Part III model. Parameters for the Part I and Part II models are fixed across all conditions.

Part I: The data were generated using a unidimensional 2PL IRT model for $J = 20$ unmodified items, with sample size $P = 1{,}000$. True latent ability scores $\boldsymbol{\theta}$ were generated from a standard normal distribution. The slope parameters were sampled from a normal (1.672, .761) distribution, and the intercepts were sampled from a normal (2.308, 1.670) distribution. The slope parameters were truncated at .5. The intercepts were also constrained such that item difficulty, defined as the negative ratio of the intercepts to slope parameters, is between -3

and 3. These were chosen to reflect the properties of the unmodified items in the ELPA21 online item bank.

Part II: The data were generated using a cross-classified random effects 2PL IRT model for either $I = 20$ or $I = 50$ dichotomous items with sample size $P = 1{,}000$. The slope parameters were drawn from normal (1.260, 0.534) distribution, truncated at .5. Note that the slope parameters are not of interest and thus fixed to the true values in the analysis. The intercept parameters follow normal (1.287, 1.230) distribution. It should be noted that actual intercept parameters differ for each replication because items were treated as random. Here, the conditions mirror the properties of the source item pool in the ELPA21 assessment.

Part III: Following ELPA21, we only consider two cut scores ($K = 2$). The two cut scores were set at $\boldsymbol{\theta}^* = (-0.5, 0.2)'$. We fixed the true intercept and regression coefficients for latent regression model as $\beta_0 = 0$ and $\beta_1 = .6$ and the error covariance matrix is specified as

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} .001 & \\ .001 & .002 \end{bmatrix}.$$

Manipulated factors are the number of raters (4), item variance (2), and rater variance (2) after accounting for the item covariate. This results in $4 \times 2 \times 2 = 16$ conditions. For $I = 20$ items, The number of raters was $R = 2, 3, 4$, and 10, the item residual variance $\sigma^2_\zeta = (.2^2, .5^2)$, and the rater variance $\sigma^2_\gamma = (.1^2, .2^2)$. Four additional conditions were considered; for $R = 2$, we increased the number of items to $I = 50$ (as in Part II), all other factors being the same. This aims to see how an increased number of items more precisely estimates the parameters in the Part II model, which is important since those estimates may impact the accuracy of Part III model parameter estimates.

Accordingly, a total of 20 conditions were used to generate the rating data from the multivariate nonlinear cross-classified random effects model (Equation 6). First, we generated "true" revised item parameters α_i. This simulates an ideal scenario where an infinite pool of raters come to a perfect agreement. These parameters are therefore the target to recover, which is an ultimate goal of this study. Again, "true" revised item parameters differ for each replication because items are treated as random. Then, rater variations were added to reflect raters' judgements. The probability scale of the ratings constrain the data to be between 0 and 1. For each condition, 100 data sets were generated. All data generation and analysis were performed in R (R Core Team, 2018).

6. Simulation Study Results

6.1. Point Estimates

We present parameter estimates from the Part II model in Table 5. Recall that we only have two conditions for the Part II model: 20 modified items for

TABLE 5.
Estimates of Part II Structural Parameters

	μ_c				σ_c^2			
I	True	Estimates	Bias	RMSE	True	Estimates	Bias	RMSE
20	1.287	1.304	.017	.278	1.513	1.452	−.061	.495
50		1.296	.009	.183		1.503	−.010	.289

Note. True = generating values; Estimates = mean of point estimates; Bias = absolute bias; RMSE = root mean squared error.

Conditions 1 through 16 and 50 modified items for Condition 17 through 20. Hence, we present only the results from Condition 1 for $I = 20$ and Condition 16 for $I = 20$. Previous studies have noted that small cluster sizes and/or large cluster variances tend to result in large error of approximation, especially for binary data (e.g., Cho & Rabe-Hesketh, 2011; Jeon et al., 2017; Joe, 2008). With $I = 20$, there is a clear downward bias of the variance. When I is increased to 50, the estimated bias and RMSE for both μ_c and σ_c^2 improved significantly.

We now turn to estimates from Part III. Results from all conditions are presented in Table 6. The error covariance matrix is omitted to save space given that they yielded both bias and RMSE of essentially 0.

Our first observation is that the latent regression intercept β_0 tended to be slightly biased downward, and the bias was generally smaller for smaller σ_ζ^2 and σ_γ^2. Specifically, within the same I and R, the estimated bias decreased as σ_ζ^2 decreased from .25 to .04 given $\sigma_\gamma^2 = .009$ or .04. Similarly, it also decreased as σ_γ^2 decreased from .04 to .009 given $\sigma_\gamma^2 = .04$ or .25. We also found a similar pattern for RMSE. Another observation is that increasing the number of raters or increasing the number of items resulted in smaller RMSE. However, increasing the number of raters did not help reduce the bias in β_0.

Second, the regression coefficient β_1 seems to be well recovered. The bias ranged from .004 to .013, and no distinct pattern was found across conditions. RMSE, on the other hand, showed clearly larger values for larger σ_ζ^2. Specifically for $I = 20$, RMSE ranged from .040 to .045 when $\sigma_\zeta^2 = .04$, and from .090 to .097 when $\sigma_\zeta^2 = .25$. In addition, similar to the trend in β_0, it appears that RMSE becomes smaller when the number of items is increased (see Conditions 1–4 vs. Condition 17–20), but the pattern is not as clear when the number of raters is increased.

Third, the variance of random item effects σ_ζ^2 showed both positive and negative bias. Although not by a large amount, the estimated bias decreased as σ_ζ^2 increased. In contrast, RMSE increased as σ_ζ^2 increased. Note that RMSE

TABLE 6.
Estimates of Part III Structural Parameters

					β_0			β_1			σ^2_ζ			σ^2_γ		
Condition	I	R	σ^2_ζ	σ^2_γ	Est.	Bias	RMSE	Est.	Bias	RMSE	Est.	Bias	RMSE	Est.	Bias	RMSE
1	20	2	.04	.009	.001	.001	.122	.593	−.007	.045	.046	.006	.018	.005	−.004	.010
2	20	2	.04	.04	.002	.002	.188	.592	−.008	.043	.046	.006	.018	.020	−.020	.043
3	20	2	.25	.009	.000	.000	.208	.596	−.004	.097	.250	.000	.083	.008	−.001	.034
4	20	2	.25	.04	−.013	−.013	.254	.598	−.002	.095	.249	−.001	.083	.022	−.018	.043
5	20	3	.04	.009	−.003	−.003	.107	.588	−.012	.044	.044	.004	.015	.007	−.002	.008
6	20	3	.04	.04	−.027	−.027	.167	.587	−.013	.045	.045	.005	.016	.028	−.012	.036
7	20	3	.25	.009	−.015	−.015	.198	.592	−.008	.096	.247	−.003	.076	.008	−.001	.009
8	20	3	.25	.04	−.049	−.049	.247	.592	−.008	.096	.248	−.002	.076	.035	−.005	.043
9	20	4	.04	.009	−.004	−.004	.101	.591	−.009	.043	.045	.005	.014	.007	−.002	.007
10	20	4	.04	.04	.003	.003	.141	.591	−.009	.040	.045	.005	.014	.034	−.006	.027
11	20	4	.25	.009	−.014	−.014	.188	.594	−.006	.094	.249	−.001	.076	.009	.000	.008
12	20	4	.25	.04	−.022	−.022	.224	.594	−.006	.093	.250	.000	.077	.040	.000	.035
13	20	10	.04	.009	−.010	−.010	.087	.592	−.008	.040	.046	.006	.015	.009	.000	.005
14	20	10	.04	.04	−.023	−.023	.111	.592	−.008	.040	.046	.006	.014	.038	−.002	.019
15	20	10	.25	.009	−.017	−.017	.170	.592	−.008	.090	.251	.001	.077	.010	.001	.005
16	20	10	.25	.04	−.036	−.036	.180	.594	−.006	.093	.251	.001	.077	.039	−.001	.018
17	50	2	.04	.009	−.003	−.003	.090	.589	−.011	.032	.044	.004	.011	.005	−.004	.009
18	50	2	.04	.04	−.031	−.031	.160	.589	−.009	.031	.044	.005	.011	.022	−.003	.040
19	50	2	.25	.009	.001	.001	.139	.590	−.010	.060	.246	−.004	.053	.007	−.002	.010
20	50	2	.25	.04	−.044	−.044	.231	.589	−.011	.061	.246	−.004	.052	.038	−.002	.068

Note. True = generating values; Est. = mean of point estimates; Bias = absolute bias; RMSE = root mean squared error; generating values for β_0 and β_1 are 0 and 0.6 across all conditions.

weights larger errors more. As expected, the number of raters does not seem to affect the results, but interestingly, no noticeable pattern was found either with respect to the number of items in terms of bias. Similar to β_0 and β_1, RMSE of σ^2_ζ reduced when the number of items increased. However, it may not be appropriate to draw a general conclusion given that only four conditions were examined for $I = 50$.

Finally, the variance of random rater effects σ^2_γ tended to be underestimated. When σ^2_γ was smaller, the bias tended to be smaller as well. The number of raters mattered in the estimation of σ^2_γ. When $R = 2$, the largest bias reached $-.020$, and when the number of raters increased to $R = 10$, it dropped to as low as $-.002$. RMSE appears to follow the same pattern. As expected, the number of items did not seem to affect the estimation of rater variance at all. Taken together, we suspect that the linchpin to successful parameter recovery hangs on the accuracy of the two important variance components: σ^2_ζ and σ^2_γ. Further studies on priors and sensitivity analysis may be needed.

6.2. SE*s*

The accuracy of postconvergence *SE*s from all replications for two conditions (4 and 17) selected for contrast are reported in Figure 1. Specifically, they compare the mean of estimated *SE*s against the Monte Carlo *SD*s of the point estimates. As the item parameters in Parts II and III are of particular interest, only these two sets are presented. Under Condition 4 (the left panel), *SE*s were recovered relatively rather poorly, and under Condition 17 (the right panel), *SE*s were recovered well. In general, we observed a close agreement between the mean of the *SE*s and the Monte Carlo *SD*s. In Part II, the two values are very close to one another in both conditions but with a slight downward bias in Condition 4. In Part III, they align well on the reference line but again with a slight downward bias especially in Condition 4.

The general underestimation of *SE*s can be explained by several reasons. First, for efficiency, we only used one imputation per MH-RM cycle in this study. As noted by Cai (2008), such a small number of imputations may lead to estimated *SE*s with a downward bias (see multiple imputation theory in Little & Rubin, 1987, for example). Second, we opted for postconvergence *SE*s and not recursive *SE*s, again for the efficiency reason. While the postconvergence approach improves the stability of estimation for point estimates, the resulting *SE*s may be prone to underestimation (e.g., Yang & Cai, 2014), though this may not always be the case in difference research contexts.

6.3. Random Effects Scoring: Revised Item Parameters

Our goal is to obtain revised item parameters for special populations. With varying levels of parameter recovery across conditions, the natural question is to

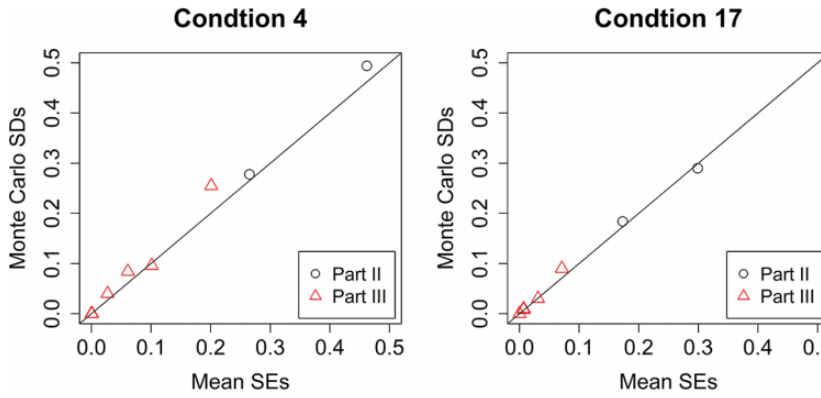


FIGURE 1. *Accuracy of Standard Errors: Conditions 4 and 17.*

what extent the differences in structural parameter recovery impact the recovery of revised item parameters, which corresponds to the EB prediction of the random item effects α_i. Pearson correlations, average bias, and average RMSE of the EB revised item parameters are presented in Table 7.

Regardless of the conditions examined, r was no less than .99, retaining the true rank-ordering almost perfectly. In terms of bias, α_i exhibit downward bias, as expected from the underestimation trends observed in the regression parameters (see Table 6). Both bias and RMSE tend to decrease with smaller variance of random rater effects σ^2_γ, suggesting rater preparation is important. It was also found that the decrease in the variance of random item effects σ^2_ζ reduced RMSE, replicating what is known in the linking/equating literature, namely, that more refined test linking is possible when the tests being linked do not differ so dramatically on core psychometric properties. Surprisingly, increasing the number of raters did not lead to the improvement of final estimates of revised item parameters, which has cost and logistics implications for operational studies.

In order to inspect the average bias, we plotted the estimated revised item parameters against "true" revised item parameters in Figure 2. Condition 1–2, 9–10, and 13–14 were specifically selected because the effects of the number of raters and the variance of random rater effects are made prominent across them. Recall that "true" revised item parameters differ across replications, and hence, each dot is for one data set. Note that dispersion about the line is due to the fact that the plot is of the estimated parameter value per replication, not of the *mean* of the estimated parameter values. Two points are noteworthy. First, the average downward bias (see Table 7) does not mean that underestimation is systematic for all replications. Second, while the effect of number of raters is not made obvious from the average bias, revised item parameters appear to be better

TABLE 7.
Recovery of Revised Item Parameters

Condition	I	R	σ^2_ζ	σ^2_γ	r	Average Bias	Average RMSE
1	20	2	.04	.009	.993	−.127	.182
2	20	2	.04	.04	.994	−.127	.227
3	20	2	.25	.009	.993	−.123	.191
4	20	2	.25	.04	.990	−.138	.240
5	20	3	.04	.009	.996	−.133	.176
6	20	3	.04	.04	.996	−.155	.229
7	20	3	.25	.009	.991	−.133	.181
8	20	3	.25	.04	.993	−.158	.233
9	20	4	.04	.009	.997	−.132	.168
10	20	4	.04	.04	.997	−.126	.190
11	20	4	.25	.009	.996	−.137	.174
12	20	4	.25	.04	.995	−.147	.205
13	20	10	.04	.009	.997	−.136	.156
14	20	10	.04	.04	.998	−.147	.175
15	20	10	.25	.009	.997	−.140	.162
16	20	10	.25	.04	.998	−.159	.186
17	50	2	.04	.009	.992	−.137	.188
18	50	2	.04	.04	.992	−.166	.245
19	50	2	.25	.009	.991	−.132	.193
20	50	2	.25	.04	.992	−.177	.280

Note. r indicates Pearson's correlation; Average Bias = average absolute bias across items; Average RMSE = average root mean squared error across items.

recovered with larger number of raters and smaller rater variance. The recovery is the best for $R = 10$ and $\sigma^2_\gamma = .009$ (Condition 13).

Finally, we examined latent proficiency recovery with the revised item parameters. In the operational context, this is the ultimate objective. As an illustration, the results from Condition 20 is displayed in Figure 3. The left panel in Figure 3 plots the Expected A Posteriori (EAP) scores obtained with the revised item parameters ("Estimated") against the simulated individual latent proficiency values, that is, true θ ("True"). The latent proficiency values are well recovered with $r = .997$. For all conditions (not displayed here), r was always greater than .99. The apparent nonlinearity stems from the shrinkage effect of EAP scores.

Additionally, we plotted the EAP scores obtained with the estimated α_is ("Estimated") against the EAP scores obtained with the "true" α_is generated in the simulation ("Benchmark"), shown on the right panel in Figure 3. The EAP scores from the revised item parameter estimates are perfectly aligned with those

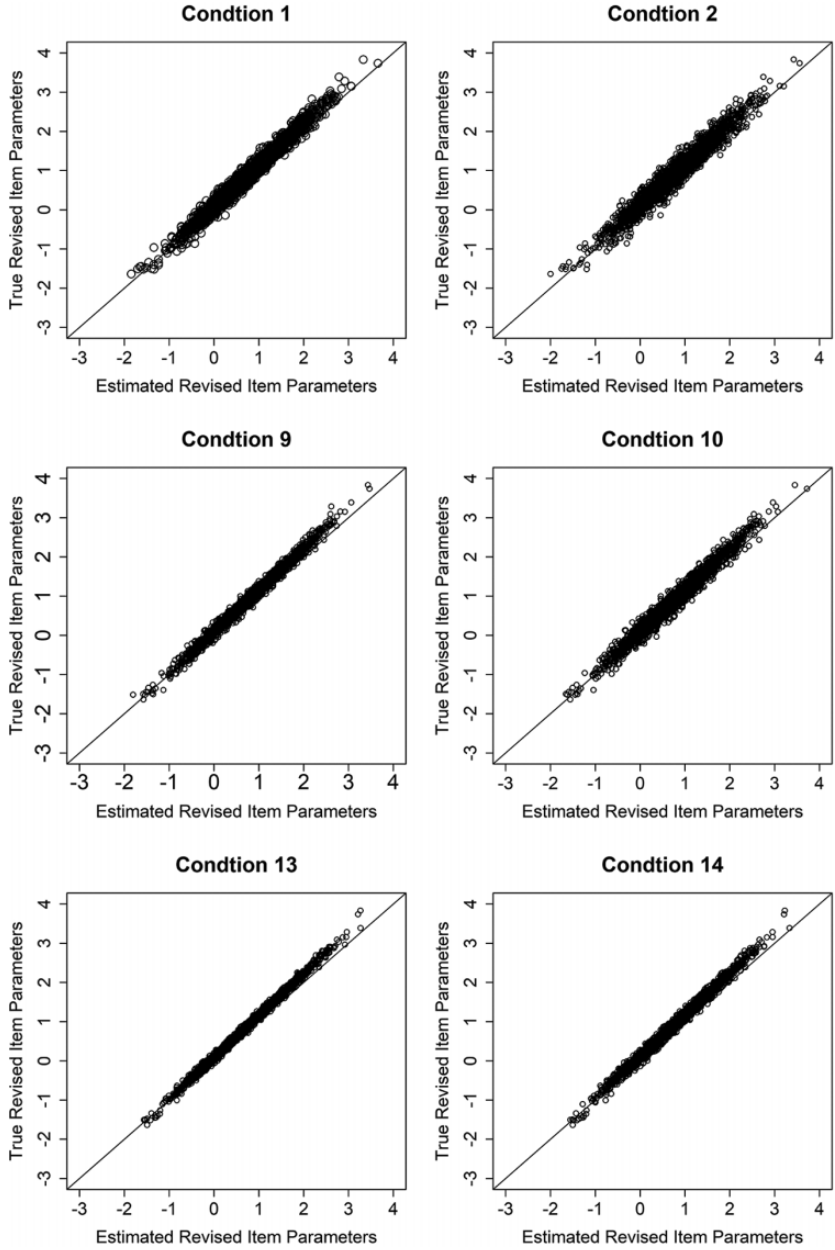


FIGURE 2. *Revised Item Parameters: Conditions 1–2, 9–10, and 13–14.* Note*: "Estimated Revised Item Parameters" is* $\hat{\alpha}$*, that is, the random item effects scores. "True Revised Item Parameters" is* α*, which simulates an ideal scenario where an infinite pool of raters come to a perfect agreement.*

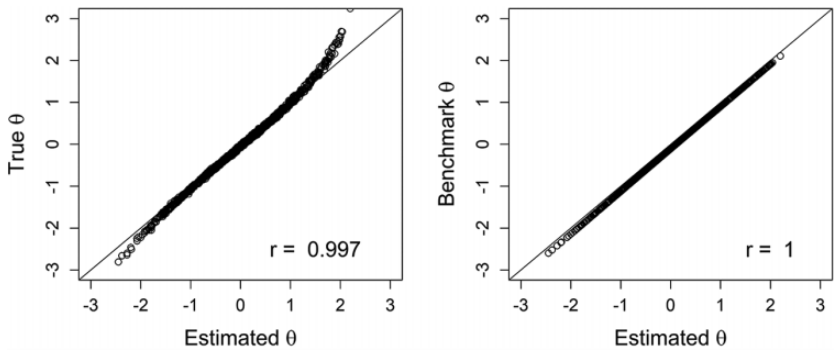


FIGURE 3. *Latent ability recovery.* Note*: "Estimated θ" is the EAP scores obtained with the estimated revised item parameters. "True θ" is the generated individual latent ability scores. "Benchmark θ" is the EAP scores obtained with the "true" revised item parameters.*

from the benchmark with $r = 1$. For all other conditions (not displayed here), r was also 1.

7. Discussions

This study developed a cross-classified random effects model that jointly utilizes item response data from a general population (prior calibrations) and judge-provided data from expert raters to revise item parameters for use in scoring modified test forms. More specifically, we introduced a three-part model, a combination of (1) a nonlinear mixed model with fixed item effects and random person effects to model responses on the unmodified portion of the original test form; (2) a generalized linear mixed model with cross-classified random effects to model responses to the modified portion of original test form; and (3) a multivariate nonlinear cross-classified random effects model with covariates to model the ratings of the target items. We use the MH-RM algorithm (Cai, 2008, 2010a, 2010b) for maximum marginal likelihood parameter estimation. Given the maximum likelihood estimates of structural parameters, EB was used to revise the item parameters.

We apply the proposed methods to empirical data from the Braille ELPA21 assessment for BLV students. We also conducted a simulation study. Given limited results, we can tentatively conclude that we may be able to obtain revised item parameters suitable for scoring individuals with a relatively small number of raters as long as raters' variance is not too large.

Our study contributes to the line of research on estimating item difficulty by judgmental methods (e.g., Farmer, 1928; Hambleton & Jirka, 2006) by providing an innovative methodological approach to moderated item calibration.

The present literature and current practices in operation take the fixed effects approach (e.g., simple average of ratings from all expert raters). This study enhances the postjudgment data analysis process. The random effects approach taken in this study brings us significant benefits. We can "borrow strength" from other parts of data obtained from other items and raters when inferring the parameters of a single item (Wainer et al., 2001).

Moderated item calibration may also have wide applicability. One notable strength of the approach is that it is useful when standard linking procedures are not feasible. Even when we have common items or common population is assumed, our approach can certainly serve as an alternative path to the current small sample equating methods (see, e.g., Albano, 2015; Kim et al., 2008; Skaggs, 2005). Previous studies have shown that when sample size is extremely small and the two test forms are not completely parallel, there is a lack of good equating solutions (Kolen & Brennan, 2004; Skaggs, 2005).

This work, of course, has some limitations stemming from the assumptions we made. For one, as a reviewer pointed out, we have not considered the departure from the normality assumption of the item random effects or the rater random effects. One could imagine that there are situations where the difficulty changes only for very specific modified items, and therefore, the normality assumption may not be appropriate. Nevertheless, we restrict the model and simulations to the normality assumption for simplicity, as it is the first step in developing the method of this kind. In future research, it would be interesting to examine the impact of nonnormaliaty on the final person-parameter estimates. Within the proposed estimation scheme, extensions to the other distributions are expected to be straightforward.

Let us now reflect on the validity of using expert judgments. Since there is no guarantee that these "experts" will judge "correctly," their involvement may seem questionable. However, as Bolsinova et al. (2017) noted, the linking data are also susceptible to bias. Importantly, our empirical application provides us with one quantitative evidence of "true" shift in item difficulty, as told by raters who work with BLV students daily. That is, the empirical data showed that the Braille items were systematically judged to be more difficult. The lack of perfect agreement among the raters leads to uncertainty. The uncertainty is quantified via our cross-classified random effects modeling. To reiterate, one of the goals of this study is to develop mechanisms to tolerate variability in expert judgments so that we can isolate systematic differences in the items' psychometric properties from the original items. Furthermore, our assumption here is that the experts should be able to utilize the information from the original assessment of the general population. This anchored design should, at minimum, make the raters more informed judges. Nevertheless, we are well aware that it is a strong assumption to rely as heavily on expert judgments, so we suggest further improvement on the data collection design might be needed. For example, raters' confidence

on their ratings can be incorporated to better estimate the revised item parameters.

Finally, we note that simply assuming that modified item parameters do not change, on the basis that the modification is done "well" is a judgment made by the assessment developer. One can debate whether this is a more evidence-driven, better judgment than the ratings provided by a group of subject matter experts who have much more familiarity and better qualified in adaptation and modification from their experience of working with BLV students on a regular basis. Perhaps one should at least check whether or not the modified item parameters have indeed not changed, and if changed, to what extent, by empowering the voices of educators.

Acknowledgments

We thank Mark Hansen, Noreen Webb, and Michael Seltzer for their thoughtful feedback. In addition, we would like to thank reviewers for their helpful comments and suggestions.

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) received no financial support for the research, authorship, and/or publication of this article.

References

Albano, A. D. (2015). A general linear method for equating with small samples. *Journal of Educational Measurement*, *52*, 55–69.

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, *6*(3), 309–325. https://doi.org/10.1177/014662168200600308

Bolsinova, M., Hoijtink, H., Vermeulen, J. A., & Bé guin, A. (2017). Using expert knowledge for test linking. *Psychological Methods*, *22*, 705–724.

Bramley, T., & Benton, T. (2017). *Comparing small-sample equating with angoff judgment for linking cut-scores on two tests* [Paper presentation]. 18th Annual Conference of the AEA Europe conference Prague, Czech Republic.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood non-linear latent structure analysis with a comprehensive measurement model* [Unpublished doctoral dissertation]. The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Bahavioral Statistics*, *35*, 307–335.

Cai, L. (2015). *flexMIRT®version 3.0: Flexible multilevel multidimensional item analysis and test scoring* [Computer software manual]. Vector Psychometric Group.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Sage.

Clayton, D., & Rasbash, J. (2011). Estimation in large cross random-effect models by data augmentation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*, 425–436.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: Method and application. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 259–273). Chapman & Hall.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement*, *53*, 3–22.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328–347.

Farmer, E. (1928). Concerning subjective judgment of difficulty. *British Journal of Psychology*, *18*, 438–442.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Lawrence Erlbaum Associates.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2017). A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, *82*, 693–716.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*, 5066–5074.

Kim, S., von Davier, A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45*, 325–342.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer-Verlag.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*, 83–102.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Lorge, I., & Diamond, L. K. (1954). The value of information to good and poor judges of item difficulty. *Educational and Psychological Measurement*, *14*, 29–33.

Lorge, I., & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement*, *12*, 554–561.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *44*, 226–233.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley & Sons.

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

Mislevy, R. J. (1993). *Linking educational assessments: Concepts, issues, methods, and prospects*. Educational Testing Service.

Monroe, S. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities* [Unpublished doctoral dissertation]. University of California, Los Angeles, CA.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve IRT model by the Metropolis-Hastings Robbins-Monroe algorithm. *Educational and Psychological Measurement*, *74*, 343–369.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). Guilford.

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*, 146–178.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software manual]. R Core Team. https://www.R-project.org/

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, *42*, 309–330.

Skrondal, A., & Rabe-Hesketh, S. (2014). *Generalized latent variable modeling*. Chapman & Hall/CRC.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 309–317). Academic Press.

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2001). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653–673). Lawrence Erlbaum.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In H. Wainer & D. Thissen (Eds.), *Test scoring* (pp. 343–387). Lawrence Erlbaum Associates.

Winter, P., Hansen, M., & McCoy, M. (2018). *Ensuring the comparability of modified tests administered to special populations* [Paper presentation]. National Council on Measurement in Education, New York, NY.

Yamamoto, K., & Mazzeo, J. (2005). Chapter 4: Item response theory scale linking in NAEP. *Journal of Educational Statistics*, *17*, 155–173.

Yang, J., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monroe algorithm. *Journal of Educational and Behavioral Statistics*, *39*, 550–582.

Authors

SEUNGWON CHUNG is an assistant professor of quantitative methods in education within the Department of Educational Psychology at the University of Minnesota, Twin Cities, 56 East River Road, Minneapolis, MN 55455; email: swchung@umn.edu. Her research interests include latent variable modeling with focus on item response theory and model fit evaluation for categorical data.

LI CAI is a professor of education and psychology in the Advanced Quantitative Methodology program within the UCLA Graduate School of Education and Information Studies, 405 Hilgard Avenue, Los Angeles, CA 90095; email: lcai@ucla.edu. His research interests include the development, integration, and evaluation of innovative latent variable models that have wide-ranging applications in assessment research in educational, psychological, and health-related domains of study.

Manuscript received February 2, 2020
First revision received October 22, 2020
Accepted December 5, 2020