Can IRT solve the missing data problem in test equating?

Bolsinova, M.; Maris, G.

DOI
10.3389/fpsyg.2015.01956

Publication date
2016

Document Version
Final published version

Published in
Frontiers in Psychology

published: 05 January 2016
doi: 10.3389/fpsyg.2015.01956

Can IRT Solve the Missing Data Problem in Test Equating?

Maria Bolsinova [1, 2]* *and Gunter Maris* [2, 3]

[1] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [2] Psychometric Research Center, Dutch National Institute for Educational Measurement (Cito), Arnhem, Netherlands, [3] Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

In this paper test equating is considered as a missing data problem. The unobserved responses of the reference population to the new test must be imputed to specify a new cutscore. The proportion of students from the reference population that would have failed the new exam and those having failed the reference exam are made approximately the same. We investigate whether item response theory (IRT) makes it possible to identify the distribution of these missing responses and the distribution of test scores from the observed data without parametric assumptions for the ability distribution. We show that while the score distribution is not fully identifiable, the uncertainty about the score distribution on the new test due to non-identifiability is very small. Moreover, ignoring the non-identifiability issue and assuming a normal distribution for ability may lead to bias in test equating, which we illustrate in simulated and empirical data examples.

Keywords: item response theory, incomplete design, marginal Rasch model, missing data, non-identifiability, test equating

OPEN ACCESS

Edited by:
Mike W.-L. Cheung,
National University of Singapore,
Singapore

Reviewed by:
Judit Antal,
The College Board, USA
Rudy Ligtvoet,
University of Cologne, Germany

****Correspondence:***
Maria Bolsinova
m.a.bolsinova@uva.nl

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *28 May 2015*
Accepted: *06 December 2015*
Published: *05 January 2016*

Citation:
Bolsinova M and Maris G (2016) Can
IRT Solve the Missing Data Problem in
Test Equating?
Front. Psychol. 6:1956.
doi: 10.3389/fpsyg.2015.01956

1. INTRODUCTION

One of the advantages of item response theory (IRT) over classical test theory is its ability to handle incomplete designs. Among the important applications in which data are missing by design is test equating, where results of different test forms must be made comparable by accounting for the two key facts. The first is that the reference and the new tests need not be of the same difficulty, and the second is that the reference and the new populations need not have the same ability distribution (Kolen and Brennan, 2004; von Davier, 2011).

Suppose, that the same students respond both to the reference and to the new test. Assume, for the sake of the argument, that both tests are scored with a number correct score. It is clear that, if both tests represent the same underlying construct, both scores are automatically equated. The need for equating scores derives from the fact that for every student we only observe the response to either the reference or the new test. That is, it derives from the fact that there is a missing data problem.

Equating procedures are methods to overcome the missing data problem. There are many different methods for score equating with some methods based on IRT and other on classical test theory. These methods are covered in detail by, for example, Kolen and Brennan (2004), von Davier (2011), von Davier et al. (2004), Holland and Dorans (2006), and Livingston (2004). Most all equating procedures are such that all students with the same score on the reference test get the same equated score on the new test. This in contrast to both the complete data case we considered above, and more modern (multiple) imputation based techniques (Rubin, 1987).

The central question we consider in this paper is whether the distribution of the missing data (marginal or conditionally on the observed data) is in principle identifiable from the observed data.

If the marginal distribution is not identifiable, neither is the conditional distribution needed to impute the missing data. Regardless of the preferred equating method, if the distribution of the missing data is not identifiable, the missing data problem can not be solved.

Suppose we take the most modest form of equating: translating the scores on the new test to a pass/fail decision (i.e., selecting a cut-score below which a student fails) consistently with the pass/fail criterion on the reference test, i.e., such the passing percentage in the reference population would be the same on the new test as it is on the reference test. To specify a new cutscore, it is sufficient to estimate the distribution of the scores of the persons from the reference population to the new test, denoted by $p(X_{+mis})$[1]. As we will show in the paper, this is not possible using an IRT model given the observed data only. Hence, solving more complicated problems of equating (obtaining a full correspondences between the scores on the two tests) is also not possible.

When IRT is used for test equating, the joint distribution of the observed data [responses of the reference population on the reference test, denoted by $p(\mathbf{X}_{obs})$] and the missing data [responses of the reference population on new test, denoted by $p(\mathbf{X}_{mis})$] is modeled by a marginal IRT model that consists of a conditional distribution of the data given a latent variable θ and a population distribution $f(\theta)$. Two elements are required to estimate the distribution of missing responses $p(\mathbf{X}_{mis})$. First, the parameters of the items from the new test and from the reference test must be placed on the common scale. Second, the ability distribution of the reference population given the observed data $f(\theta \mid \mathbf{X}_{obs})$ must be estimated. In this paper, we have assumed that the tests are well connected through a linking design[2] and the IRT model is correctly specified and, therefore, the first element of equating is fully satisfied. We have focused on the second element, which is usually ignored in test equating practice. The problem is that the full distribution of ability $f(\theta)$ is not identifiable, as has been shown by Cressie and Holland (1983). Consequently, as we show in this paper, the distribution $p(X_{+mis})$ is also not identified from the observed data only. This issue is usually ignored in test equating practice, and instead a parametric distribution, usually a normal distribution, is assumed for $f(\theta)$. This assumption is not guaranteed to hold in practice, therefore it is important to consider to what extent the problem of inferring the distribution of missing responses can be solved without extra distributional assumptions.

We will discuss the problem of non-identifiability of $p(X_{+mis})$ using the marginal Rasch model (RM) for dichotomous data, which has only one parameter in the conditional model (Rasch, 1960), as an example. The RM is chosen here for convenience; the identifiability issues are present at the level of the marginal model and are therefore not affected by the choice of a particular parametric conditional model.

In this study we investigate the extent to which the unavoidable uncertainty about the score distribution $p(X_{+mis})$ that comes from non-identifiability is problematic in practice. The main purpose of this study is not to introduce a new method for test equating, but to highlight a fundamental property of marginal IRT models. This property is that in IRT equating the score distribution $p(X_{+mis})$ can not be identified without making extra assumptions about the parametric shape of the ability distribution, and the practical consequences of ignoring this property.

2. WHY IRT CANNOT SOLVE THE MISSING DATA PROBLEM

In this section we describe a simple model for test equating that tries (unsuccessfully) to predict missing responses from the observed data without additional distributional assumptions. The marginal RM is:

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \int_{\mathbb{R}} \prod_i \frac{\exp(x_i(\theta - \delta_i))}{1 + \exp(\theta - \delta_i)} f(\theta)\, d\theta, \tag{1}$$

where $\mathbf{x}$ is a vector of dichotomous responses with $x_i = 1$ if item i is answered correctly and 0 otherwise; δ_i is the difficulty parameter of item i. There is assumed to be a population distribution $f(\theta)$; however, its parametric shape is not known.

Following Cressie and Holland (1983), the marginal RM in Equation (1) can be re-written as

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \prod_i (\exp(-\delta_i))^{x_i} \int_{\mathbb{R}} (\exp(\theta))^{x_+} \prod_i \frac{1}{1 + \exp(\theta - \delta_i)} f(\theta)\, d\theta, \tag{2}$$

where x_+ is the number of items answered correctly. It can be seen that

$$f(\theta \mid \mathbf{X}_{obs} = \mathbf{0}) \propto \prod_i \frac{1}{1 + \exp(\theta - \delta_i)} f(\theta), \tag{3}$$

which is the posterior distribution of ability given that the responses to all items are incorrect. Therefore,

$$p(\mathbf{X}_{obs} = \mathbf{x}) \propto \prod_i (\exp(-\delta_i))^{x_i} E((\exp(\Theta))^{x_+} \mid \mathbf{X}_{obs} = \mathbf{0}). \tag{4}$$

To make $p(\mathbf{X}_{obs} = \mathbf{x})$ a proper density, a normalizing constant should be added. A convenient parameterisation of the marginal RM (Maris et al., 2015) is:

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \frac{\prod_i b_i^{x_i} \lambda_{x_+}}{\sum_{s=0}^{n} \gamma_s(\mathbf{b}) \lambda_s}, \tag{5}$$

where $\mathbf{b} = \{b_1, b_2, \ldots, b_n\}$ is a vector of item parameters that are transformations of difficulty parameters: $b_i = \exp(-\delta_i)$; $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \ldots, \lambda_n\}$ is a vector of population parameters, and $\gamma_t(\mathbf{b})$ denotes a t-th order elementary symmetric polynomial (Verhelst

[1] For simplicity, we considered a situation in which the new and the reference test do not have any common items. In the general case, the missing data are responses to the items that belong to the new test but not the reference test.

[2] For a review of different linking designs see, for example, Angoff (1971), Wright and Stone (1979), Lord (1980), Petersen et al. (1989), and Kolen and Brennan (2004). Some of these linking designs are presented in Appendix D.

et al., 1984). The denominator ensures that the distribution integrates to 1. The model in Equation (5) is a marginal Rasch model if and only if $\boldsymbol{\lambda}$ is a sequence of moments of a distribution. This imposes a set of inequality constraints on the parameters (Shohat and Tamarkin, 1943):

$$\det \begin{bmatrix} \lambda_0 & \lambda_1 & \dots & \lambda_m \\ \lambda_1 & \lambda_2 & \dots & \lambda_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m & \lambda_{m+1} & \dots & \lambda_{2m} \end{bmatrix} \geq 0, m = 0, 1, 2, \dots \quad (6)$$

and

$$\det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{m+1} \\ \lambda_2 & \lambda_3 & \dots & \lambda_{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m+1} & \lambda_{m+2} & \dots & \lambda_{2m+1} \end{bmatrix} \geq 0, m = 0, 1, 2, \dots \quad (7)$$

The extended Rasch model [ERM] (Tjur, 1982; Cressie and Holland, 1983; Maris et al., 2015) does not have these restrictions.

We now apply the ERM to test equating. Let us consider the joint density of the response vectors $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$:

$$p(\mathbf{X}_{obs} = \mathbf{x}, \mathbf{X}_{mis} = \mathbf{x}^*) = \frac{\prod_{i=1}^{n} b_i^{x_i} \prod_{j=1}^{m} d_j^{x_j^*} \eta_{x_+ + x_+^*}}{\sum_{t=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d}) \eta_t}, \quad (8)$$

where $\mathbf{d} = \{d_1, \dots, d_m\}$ are the parameters of the items in the new test (analogous to $\mathbf{b}$) and $\boldsymbol{\eta} = \{\eta_0, \eta_1, \dots, \eta_{n+m}\}$ is a vector of $(n + m + 1)$ population parameters corresponding to a combined test consisting of the items from both the reference and the new exams. It can be derived that the marginal distribution of the scores of the reference population on the new test is (see Appendix A, for details):

$$Pr(X_{+mis} \leq T) = \sum_{t=0}^{T} p(X_{+mis} = t) = \sum_{t=0}^{T} \frac{\gamma_t(\mathbf{d}) \sum_{s=0}^{n} \gamma_s(\mathbf{b}) \eta_{s+t}}{\sum_{u=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d}) \eta_u}. \quad (9)$$

The expression for this distribution contains parameters $\boldsymbol{\eta}$, whereas the density of the observed data contains parameters $\boldsymbol{\lambda}$. The parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ are related to each other as follows (see Appendix A, for details):

$$\lambda_s = \sum_{t=0}^{m} \gamma_t(\mathbf{d}) \eta_{t+s}, \forall s \in [0, n]. \quad (10)$$

The parameters $\boldsymbol{\lambda}$ are identified from the data (up to a multiplicative constant), whereas parameters $\boldsymbol{\eta}$ are not; this is because in this system of $(n{+}1)$ Equations (4) there are $(n{+}m{+}1)$ unknowns. Therefore, having observed only data $\mathbf{X}_{obs}$, we cannot make direct inferences about the distribution of X_{+mis}. Hence, IRT cannot solve the missing data problem.

3. WHAT IRT ALLOWS US TO INFER ABOUT THE DISTRIBUTION OF MISSING RESPONSES

The conclusion at the end of the previous section does not mean that we do not know anything about the parameters $\boldsymbol{\eta}$ or the score distribution. The relations between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ impose restrictions on the values that $\boldsymbol{\eta}$ can take, and therefore, on the score distribution. Before considering what is and is not known about the score distribution $p(X_{+mis})$, we should discussed some additional constraints for parameters $\boldsymbol{\eta}$.

Along with the restriction given by the relations with the identified parameters (10), there are other restrictions that the parameters $\boldsymbol{\eta}$ must satisfy in order to be parameters of the ERM. First, they must be positive to ensure that all probabilities in Equation (9) are positive. To derive a second constraint, consider the probability of answering item i correctly given the rest score on the test:

$$\begin{aligned} \Pr(X_i = 1 \mid X_{+obs}^{(i)} + X_{+mis} = s) &= \frac{\Pr(X_i = 1, X_{+obs}^{(i)} + X_{+mis} = s)}{\Pr(X_{+obs}^{(i)} + X_{+mis} = s)} \\ &= \frac{b_i \gamma_s(\mathbf{b}^{(i)}, \mathbf{d}) \eta_{s+1}}{b_i \gamma_s(\mathbf{b}^{(i)}, \mathbf{d}) \eta_{s+1} + \gamma_s(\mathbf{b}^{(i)}) \eta_s} = \frac{b_i \frac{\eta_{s+1}}{\eta_s}}{1 + b_i \frac{\eta_{s+1}}{\eta_s}}, \end{aligned} \quad (11)$$

where $\mathbf{b}^{(i)}$ denotes a vector of item parameters of all items in the reference test except item i, and $X_{+obs}^{(i)}$ is the sum score on these items; that is, the rest score. From the measurement perspective, this probability should increase when s increases (Junker, 1993; Junker and Sijtsma, 2000). This ensures that all item-rest correlations are positive, so that it makes sense to score the particular set of items together as one test. For this to be true, the ratios $\frac{\eta_{s+1}}{\eta_s}$ must form a monotonically increasing sequence. The inequality constraint

$$\frac{\eta_1}{\eta_0} \leq \frac{\eta_2}{\eta_1} \leq \frac{\eta_3}{\eta_2} \leq \dots \leq \frac{\eta_{n+m}}{\eta_{n+m-1}} \quad (12)$$

can be specifies as a part of in the prior distribution of the population parameters (see Appendix E, for details).

An alternative motivation for using the constraints in Equation (12) is that they follow from an important feature of the marginal RM, namely that

$$\frac{\eta_{s+2}}{\eta_s} - \left(\frac{\eta_{s+1}}{\eta_s}\right)^2 \quad (13)$$

is the (posterior) variance of $\exp(\theta)$ of a person with a score of s (Maris et al., 2015). The monotonicity constraints in Equation (12) follow from non-negativity of variance. Therefore, the constraints in Equation (12) are necessary but not sufficient for the parameters to satisfy the moment constraints of the marginal RM. Hence, the model we are using for equating is an ERM with the monotonicity constraints. As will be shown in the next subsection this restriction enables to reduce the uncertainty about the score distribution on the new test.

3.1. A Simple Case: $m = 1$

In this subsection we derive the uncertainty about the marginal probability of answering a new item correctly, given the observed responses to n items. Let us consider the simplest case in which the number of items in the new test is equal to one ($m = 1$). Because we are ignoring the effect of sampling variability on the uncertainty, we consider all identifiable parameters ($\mathbf{b}$ and $\boldsymbol{\lambda}$) known.

Let $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \ldots, \lambda_n\}$ denote the set of identifiable population parameters; $\boldsymbol{\eta} = \{\eta_0, \eta_1, \eta_2, \ldots, \eta_{n+1}\}$ the set parameters for the combined test; and d the item parameter of the new item. The relations between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ form a system of linear equations:

$$\begin{pmatrix}\lambda_0\\ \lambda_1\\ \vdots\\ \lambda_n\end{pmatrix} = \begin{pmatrix}1 & \mathrm{d} & \ldots & 0 & 0\\ 0 & 1 & \ldots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \ldots & 1 & \mathrm{d}\end{pmatrix}\begin{pmatrix}\eta_0\\ \eta_1\\ \vdots\\ \eta_n\\ \eta_{n+1}\end{pmatrix}. \tag{14}$$

This system of $n + 1$ equations does not have a unique solution because the number of unknowns ($n + 2$) is larger than the number of equations. The general solution is:

$$\begin{pmatrix}\eta_0\\ \eta_1\\ \vdots\\ \eta_{n+1}\end{pmatrix} = \begin{pmatrix} k\\ \frac{\lambda_0}{d} - \frac{k}{d}\\ \vdots\\ \sum_{t=0}^{n}\frac{(-1)^{n-t}\lambda_t}{d^{n+1-t}} + (-1)^{n+1}\frac{k}{d^{n+1}}\end{pmatrix}, \tag{15}$$

where k is a parameter that captures all uncertainty about $\boldsymbol{\eta}$, such that the unique solution to the system of equations can be computed when k is known. This parameter is not completely free because $\boldsymbol{\eta}$ must satisfy the set of inequalities:

$$\begin{cases}\eta_s > 0, \forall s \in [0 : (n+1)],\\ \frac{\eta_{s+1}}{\eta_s} \geq \frac{\eta_s}{\eta_{s-1}}, \forall s \in [1 : n].\end{cases} \tag{16}$$

We are interested in the probability of answering the new item correctly, which can be written as a function of k:

$$\Pr(X_{mis} = 1) = \pi_+(k) = \frac{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1}}{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1} + \sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_t}. \tag{17}$$

Using the solutions of the system of equations, one can derive (for details, see Appendix B):

$$\pi_+(k) = 1 - \frac{k\sum_{t=0}^{n}\frac{(-1)^{t-1}\gamma_t(\mathbf{b})}{d^t}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t} + \frac{\sum_{t=1}^{n}\sum_{s=0}^{t-1}\frac{(-1)^{t-s}\gamma_t(\mathbf{b})\lambda_t}{d^{t-s}}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t}. \tag{18}$$

This expression is linear in k. Therefore, the uncertainty about the probability of answering the new item correctly depends on the difference between the maximum and the minimum of k. The upper and the lower bounds for k can be derived from the inequalities for $\boldsymbol{\eta}$ in Equation (16).

From the non-negativity of the parameters $\boldsymbol{\eta}$ (the first set of inequalities in Equation 16), we have (see Appendix B, for details):

$$\max(0, \max_{u=1}^{\lfloor\frac{n+1}{2}\rfloor}\sum_{t=0}^{2u-1}(-1)^t\lambda_t d^t) < k < \min_{u=0}^{\lfloor\frac{n}{2}\rfloor}\sum_{t=0}^{2u}(-1)^t\lambda_t d^t. \tag{19}$$

Moreover, the second set of inequalities in Equation (16) leads to (see Appendix B):

$$\max_{u=0}^{\lfloor\frac{n-1}{2}\rfloor}\left(\frac{\lambda_{2u}^2 d^{2u}}{\lambda_{2u+1}d + \lambda_{2u}} + \sum_{t=0}^{2u-1}(-1)^t\lambda_t d^t\right) \leq k \leq \min_{u=1}^{\lfloor\frac{n}{2}\rfloor}\left(\sum_{t=0}^{2u-2}(-1)^t\lambda_t d^t - \frac{\lambda_{2u-1}^2 d^{2u-1}}{\lambda_{2u}d + \lambda_{2u-1}}\right). \tag{20}$$

Equations (19) and (20) together provide the lower and the upper bounds for k.

Next, we present a small example to show how the bounds on k change and what the uncertainty about the marginal probability of a correct response to the new item under the ERM is for different values of n. The item parameter d of this item varied from $\exp(-2)$ to $\exp(2)$, corresponding to the difficulty parameter varying from 2 to –2. We show how large the uncertainty is when only the non-negativity constraints are used, and when both the non-negativity and monotonicity constraints are used.

A data set with responses of persons sampled from a population with an ability distribution $\mathcal{N}(0, 1)$ to a test of six items with difficulties sampled from $\ln(b_i) \sim \mathcal{N}(0, 1)$ was simulated. First, only three items were taken into account, then four items, five items and, finally, all six items. We considered the identifiable parameters $\mathbf{b}$ and $\boldsymbol{\lambda}$ known in order to evaluate the uncertainty about π_+ coming only from the non-identifiability of $\boldsymbol{\eta}$. The identifiable parameters were fixed at their EAP-estimates obtained with a Gibbs sampler for the ERM (Maris et al., 2015), see **Table 1**.

The possible range of values for the free parameter k, and therefore for the probability of interest π_+ (given the fixed values of $\mathbf{b}$ and $\boldsymbol{\lambda}$) was evaluated for different values of the difficulty of the new item. **Figure 1** shows the possible ranges of values for the probability of answering the new item correctly when only the constraints in Equation (19) were used (in gray) and when the constraints in Equations (19) and (20) were used (in black).

The uncertainty about π_+ decreases when n increases. For $n = 3$, the difference between the maximum and the minimum of π_+ is for some d larger than 0.15 when only

TABLE 1 | Item and population parameters used in the illustrative example.

n	**b**	**λ**
3	{1.00, 0.58, 0.41}	{1.00, 0.80, 1.16, 3.25}
4	{8.90, 1.00, 0.58, 0.41}	{1.00, 0.52, 0.45, 0.68, 1.99}
5	{8.91, 1.12, 1.00, 0.58, 0.41}	{1.00, 0.42, 0.29, 0.32, 0.60, 2.01}
6	{8.86, 1.12, 1.00, 0.85, 0.58, 0.41}	{1.00, 0.36, 0.22, 0.19, 0.27, 0.63, 2.43}

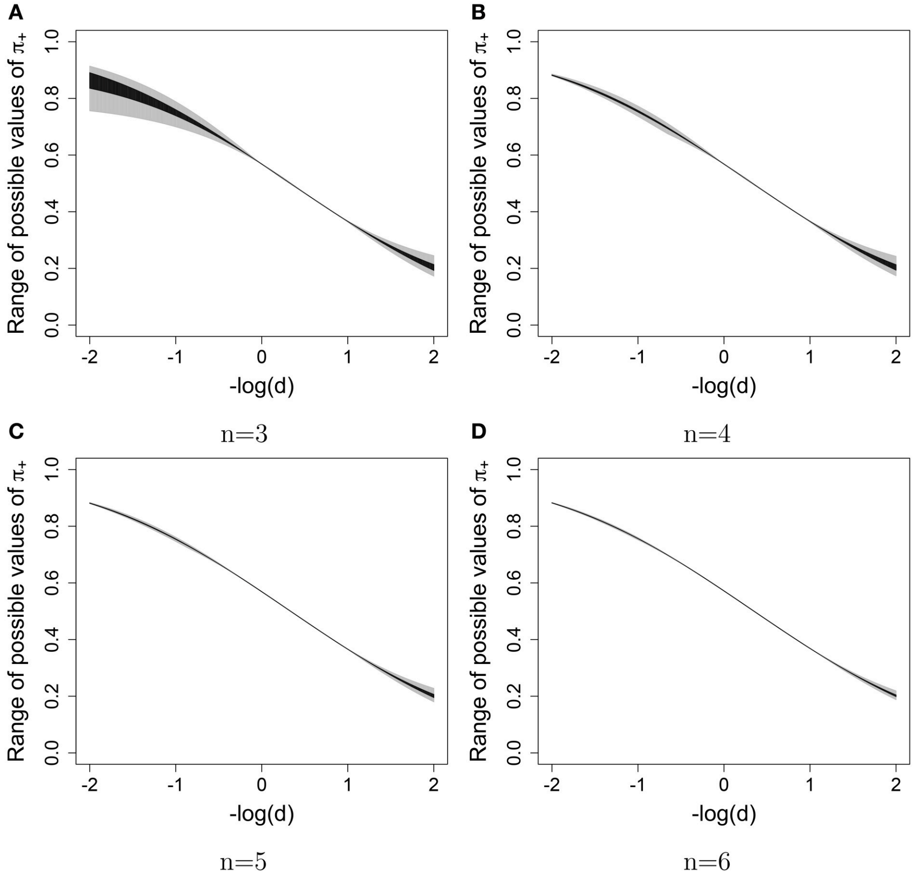

FIGURE 1 | Uncertainty about the marginal probability of answering a new item correctly (gray—without monotonicity constraints, black—with monotonicity constraints) given the difficulty of the new item (on the x-axis). (A) $n = 3$, **(B)** $n = 4$, **(C)** $n = 5$, **(D)** $n = 6$.

the constraints in Equation (19) were used and larger than 0.05 when all the constraints were used; however, when $n = 6$, the maximum discrepancy is 0.03 and 0.006 when only non-negativity constraints and all the constraints were used, respectively. Moreover, the uncertainty about π_+ for the items with the difficulty parameter close to the items that have been answered is already very small if $n = 3$. In general, the uncertainty is larger for items with extreme difficulty[3].

We have used this small example to explicitly show that it is not possible to compute the marginal probability of answering the new item correctly. However, there uncertainty about this probability is not large.

It is difficult to extend the analytic solution described in this section to realistic settings, with *n* and *m* being usual test lengths, because of the accumulation of error while computing the bounds for *k*. Therefore, below we present a simulation-based approach to the problem. Appendix C presents a proof of the fact that a simulation based approach is justifiable.

3.2. Simulated Examples

This subsection provides two simulated examples to illustrate the following:

1. the size of the uncertainty about the score distribution and which part of it is due to the non-identifiability of the parameters;
2. the practical consequences of ignoring the issue of non-identifiability of $f(\theta)$ when the true ability distribution is not normal.

In the first example, the data were simulated according to the non-equivalent group design with three linking groups. Each group consisted of 500 persons who gave responses to 15 items from the new test and 15 items from the reference test. The relevant equating designs are described in the Appendix D. The following parameters were used: $n = m = 60$, $N = M = 5000$. Responses were simulated according to the simple RM, with person parameters sampled from $\mathcal{N}(0, 1)$ for the reference population, $\mathcal{N}(0.5, 0.8^2)$ for the new population and $\mathcal{N}(-0.5, 2^2)$, $\mathcal{N}(-0.2, 2^2)$, $\mathcal{N}(-0.1, 2^2)$ for the three linking groups[4]. The item difficulties $(-\ln b_i)$ were sampled from a standard normal distribution.

First, the data augmented Gibbs sampler for the ERM with monotonicity constraints was used to estimate the total

[3] The graphs are not symmetric because the difficulties of the items in the reference test $(-\ln \mathbf{b})$ were also not symmetric around zero.

[4] These values could be seen as matching empirical practice in the sense that the persons in the linking groups perform worse than in the examination conditions and are more heterogeneous.

uncertainty about the score distribution. Second, to eliminate the uncertainty coming from the sampling variability, the new data were simulated with the same parameters but larger sample sizes ($N = M = 1,000,000$) and the algorithm was used with all the item parameters fixed at their true values. The posterior variance of the score distribution that remained was almost entirely due to the non-identifiability of the population parameters. **Figure 2** presents the widths of the 95% credibility intervals of $\Pr(X_{+mis} \leq T), \forall T \in [0 : m]$ based on 50,000 draws from the posterior distribution after 10,000 iterations of burn-in. With a large N and fixed item parameters, the uncertainty about the score distribution becomes very small, not exceeding 0.002 on the probability scale.

In the second example, we compared the results of test equating using a marginal RM assuming a normal distribution of ability in the population with the results of test equating using the ERM without the normality assumption. For the R-code of the analysis and the output, see Supplementary Material. The data with different distributions of ability in the reference population were simulated. To show what happens if normality is violated, we used skew-normal distribution for ability (Azzalini, 2005). The parameters of the skew-normal distribution were chosen such that the mean was equal to 0, variance was equal to 1, and skewness was varied $\gamma = -0.25, -0.5, -0.75$. These distributions can be seen in **Figure 3** (dotted lines) next to the standard normal distribution (solid line).

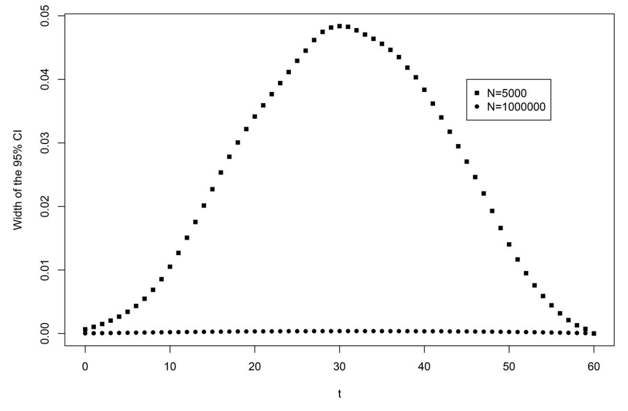

FIGURE 2 | Uncertainty about the score distribution of the reference population on the new test.

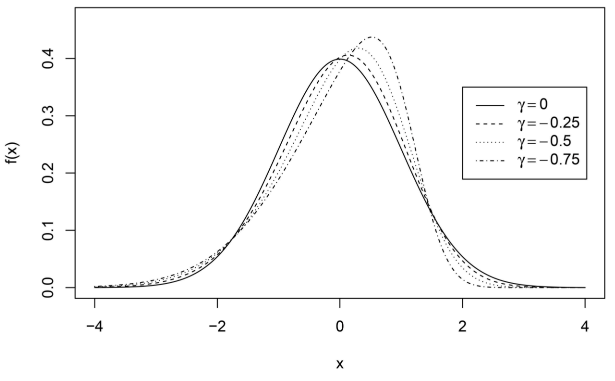

FIGURE 3 | Specification of the skewed ability distributions.

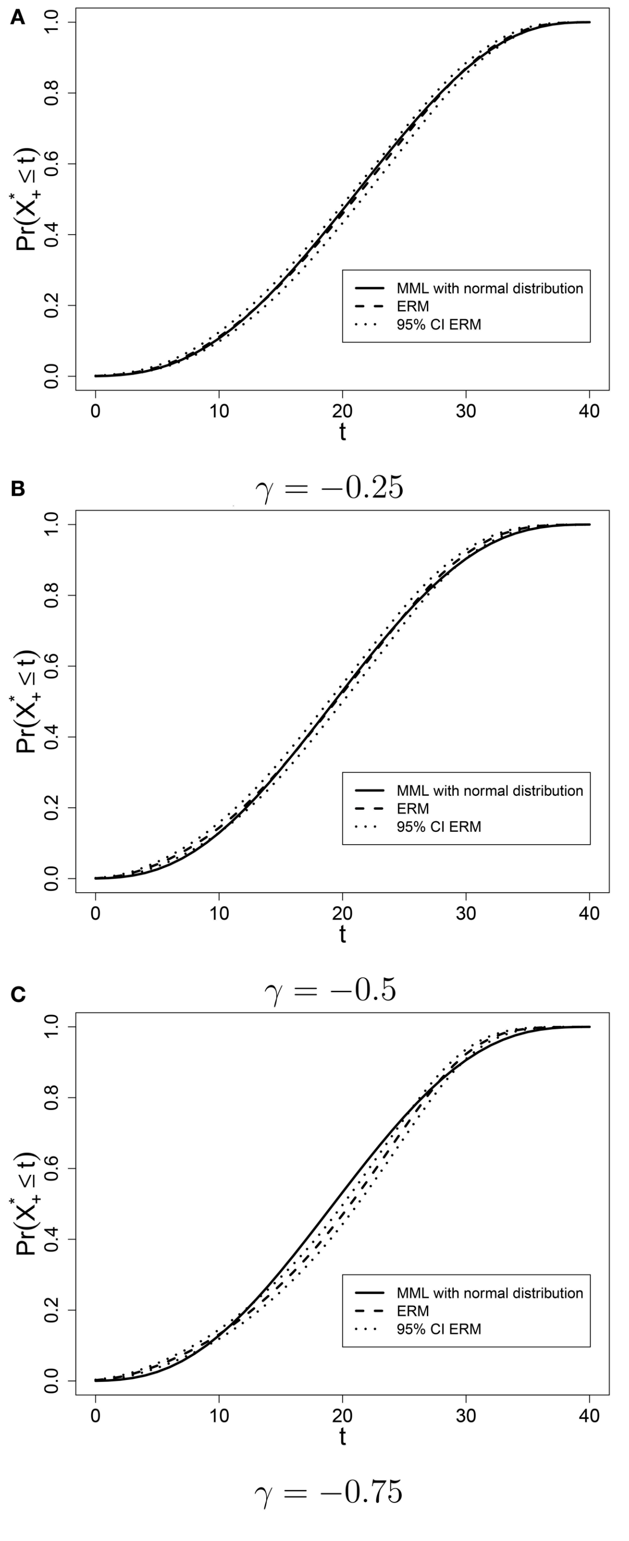

FIGURE 4 | Estimated score distributions with MML and ERM when the ability distribution in the reference population is skewed. (A) $\gamma = -0.25$; **(B)** $\gamma = -0.5$; **(C)** $\gamma = -0.75$.

For each of the three degrees of skewness, we simulated the data of 5000 persons from both the reference and the new populations taking the tests, which consisted of 40 items each, connected through three linking groups consisting of 500 persons responding to 20 items (10 from the reference test and 10 from the new test). For the new population and the three linking groups, person parameters were sampled from a normal distribution [$\mathcal{N}(0.5, 0.9^2)$, $\mathcal{N}(-0.5, 2^2)$, $\mathcal{N}(-0.2, 2^2)$, $\mathcal{N}(-0.1, 2^2)$, respectively]. Item difficulties were sampled from $\mathcal{N}(0, 1)$. The data were simulated according to a RM.

The score distribution $\Pr(X_{+mis} \leq T)$ was estimated with marginal maximum likelihood (MML) assuming a normal distribution and with the Gibbs Sampler for the ERM. The ERM score distribution together with the 95% credibility intervals of $\Pr(X_{+mis} \leq T)$ based on 50,000 draws from the posterior distribution (after 10,000 iterations of burn-in) are presented in **Figure 4**, together with the MML-estimate of the score distribution. The more skewed the ability distribution is, the greater the difference between equating results for the MML and ERM approaches. When $\gamma = -0.25$, the MML-estimate does not fall outside of the 95% credibility interval obtained with the ERM. When $\gamma = -0.5$, the estimate based on the normality assumption is outside the credible interval for low and high scores, but within the interval for the middle range of the scores. Finally, when $\gamma = -0.75$, the MML-estimate is also outside the credible bound in the middle range of test scores. This is the range of scores within which the cutscore is usually placed, which means that different score distributions are likely to result in different cutscores. This has consequences for the pass/fail decision for hundreds of students.

4. EMPIRICAL DATA EXAMPLE

Using an empirical example we show the consequences of ignoring the problem of non-identifiability of $f(\theta)$ and assuming a normal distribution. We do this by comparing the estimated score distributions with and without the normality assumption.

4.1. Method and Data

We analyzed data from the paper-and-pencil French language test for preparatory middle-level applied secondary education from examinations in 2011 and 2012. The sample sizes were 5518 for the reference exam and 5606 for the new exam. Both tests consisted of 41 items, but only dichotomous items were selected for analysis (35 and 34 in the reference and the new exams, respectively). The tests were linked through seven linking groups (with sample sizes ranging from 337 to 460) that responded to some items from either the reference test or the new test and some external anchor items (14 per group). The equating design is shown in **Figure 5**. There were 30 items from the reference test and 25 items from the new test answered by the linking groups. The items taken by the linking groups had been also answered by students in 2008.

First, the parameters of the ERM were estimated using the data augmented Gibbs sampler (see Appendix E). The algorithm was run for 60,000 iterations, of which the first 10,000 were discarded as a burn-in. The score distribution of the reference population on the new test was calculated at every iteration of the algorithm. Second, the marginal Rasch model with the normal distribution was fitted to the data and the MML-estimate of the score distribution was obtained. See Supplementary Material, for the data, software code of the analysis and the output.

4.2. Results

Figure 6 shows the posterior mean of the score distribution estimated with the ERM (together with the 95% credible interval) and the MML-estimate of the the score distribution. The estimated score distributions differ and the MML-estimate is outside of the credible interval at the lower and the higher scores. The posterior mean is also different from the MML-estimate in

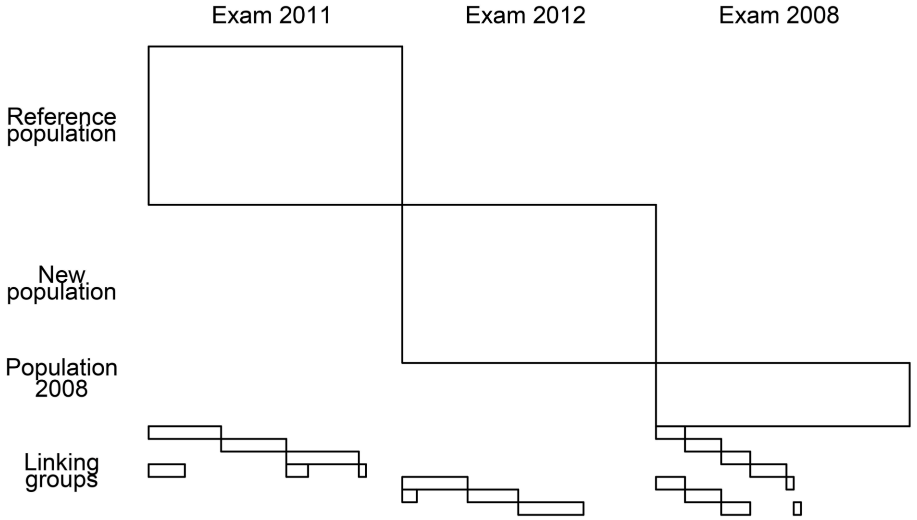

FIGURE 5 | Equating design.

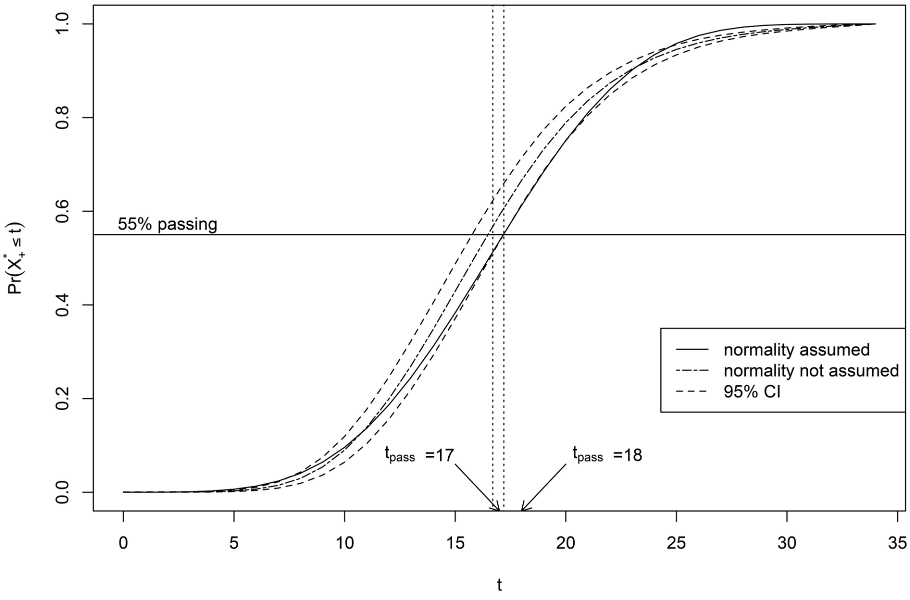

FIGURE 6 | Score distribution of the reference population on the new test: posterior mean for the ERM (dashed line) and the MML-estimate based on the assumption of the normal distribution (solid line).

the middle range of scores, which could have consequences for establishing the new cutscore t_{new}. For example, if the desired proportion of persons from the reference population failing the new test was 55%, then the MML procedure would result in a cutscore of 17, whereas the ERM procedure would result in a cutscore of 18 as illustrated in **Figure 6**. The consequence of this would be that 476 students would have passed the test if a normal distribution were assumed, but would have failed if the ERM were used.

5. DISCUSSION

Using a simple case, we have shown that, without the assumption of a parametric distribution, the score distribution on the new test is not identified. Knowing the difficulty parameter of the new item is not enough to predict the proportion of correct responses to this item in the population, after observing the responses to a finite set of items. When the number of items observed increases, the uncertainty about the score distribution decreases. This uncertainty tends to zero with *n* going to infinity, but is always there. Hence, IRT cannot, strictly speaking, solve the missing data problem, since it does not allow us to impute the unobserved responses of the reference population on the new test. We have investigated the degree of uncertainty about the score distribution in realistic applications. With realistic test lengths, the uncertainty coming from non-identifiability of population parameters is small enough to be ignored for practical purposes. Therefore, test equating can be done effectively without the not-fully-testable assumption of a particular parametric shape of the ability distribution, despite the non-identifiability issue.

The theoretical importance of this paper is that it has shown what one can and cannot do with respect to test equating using IRT based only on the observed data without the assumption of a parametric shape of the distribution. Although we have used the marginal RM for illustration, the issue of non-identifiability that is discussed holds in more general marginal IRT models, since the problem of the ability distribution not being identified will not go away if more parameters are added to the conditional model.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01956

REFERENCES

Angoff, W. (1971). "Scales, norms and equivalent scores," in *Educational measurement,* ed R. Thorndike (Washington, DC: American Council of Education), 508–600.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scand. J. Stat.* 32, 159–188. doi: 10.1111/j.1467-9469.2005.00426.x

Casella, G., and George, E. (1992). Explaining the Gibbs sampler. *Am. Stat.* 43, 167–174.

Cressie, N., and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika* 48, 129–41. doi: 10.1007/BF02314681

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Holland, P. W., and Dorans, N. J. (2006). "Linking and equating," in *Educational measurement, 4th Edn.,* ed R. L. Brennan (Westport: Praeger), 189–220.

Holland, P. W. (1990). The Dutch identity: a new tool for the study of item response models. *Psychometrika* 55, 5–18. doi: 10.1007/BF02294739

Junker, B. W., and Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Appl. Psychol. Meas.* 24, 65–81. doi: 10.1177/01466216000241004

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Ann. Stat.* 21, 1359–1378. doi: 10.1214/aos/1176349262

Kolen, M. J., and Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York, NY: Springer.

Livingston, S. A. (2004). *Equating Test Scores (without IRT).* Princeton: Educational Testing Service.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Maris, G., Bechger, T., and San Martin, E. (2015). A Gibbs sampler for the (Extended) marginal Rasch model. *Psychometrika* 80, 859–879. doi: 10.1007/s11336-015-9479-4

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989). "Scaling, norming and equating," in *Educational Measurement,* ed R. L. Linn (New York, NY: American Council of Education and Macmillan), 221–262.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Expanded edition, 1980, Chicago, IL: The University of Chicago Press; Copenhagen: the danish institute of educational research.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Survey.* New York, NY: Wiley.

Shohat, J. H. and Tamarkin, J. D. (1943). *The Problem of Moments.* New York, NY: American Mathematics Society.

Tanner, M., and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–540. doi: 10.1080/01621459.1987.10478458

Tjur, T. (1982). A connection between Rasch's item analysis and a multiplicative poisson model. *Scand. J. Stat.* 9, 23–30.

Verhelst, N. D., Glass, C. A. W., and van der Sluis, A. (1984). Estimation problems in the Rasch model: the basic symmetric functions. *Comput. Stat. Q.* 1, 245–262.

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The Kernel Method of Test Equating.* New York, NY: Springer.

von Davier, A. A. (2011). *Statistical Models for Test Equating, Scaling, and Linking.* New York, NY: Springer.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design.* Chicago, IL: MESA Press University of Chicago.

Zeger, K., and Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Stat. Assoc.* 86, 79–86. doi: 10.1080/01621459.1991.10475006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

From Equation (8), we can derive the joint distribution of the scores on the reference test and the new test:

$$\begin{aligned} p(X_{+obs} = x_+, X_{+mis} = x_+^*) &= \frac{p(\mathbf{X}_{obs} = \mathbf{x}, \mathbf{X}_{mis} = \mathbf{x}^*)}{p(\mathbf{X}_{obs} = \mathbf{x}, \mathbf{X}_{mis} = \mathbf{x}^* \mid X_{+obs} = x_+, X_{+mis} = x_+^*)} \\ &= p(\mathbf{X}_{obs} = \mathbf{x}, \mathbf{X}_{mis} = \mathbf{x}^*) \Big/ \frac{\prod_{i=1}^n b_i^{x_i} \prod_{j=1}^m d_j^{x_j^*}}{\gamma_{x_+}(\mathbf{b})\gamma_{x_+^*}(\mathbf{d})} \\ &= \frac{\gamma_{x_+}(\mathbf{b})\gamma_{x_+^*}(\mathbf{d})\eta_{x_+ + x_+^*}}{\sum_{t=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d})\eta_t}. \end{aligned} \tag{A1}$$

The marginal probability of obtaining a particular score on the new exam is then:

$$\begin{aligned} p(X_{+mis} = x_+^*) &= \sum_{s=0}^{n} p(X_{+obs} = s, X_{+mis} = x_+^*) \\ &= \frac{\gamma_{x_+^*}(\mathbf{d}) \sum\limits_{s=0}^{n} \gamma_s(\mathbf{b})\eta_{s+x_+^*}}{\sum_{u=0}^{n+m} \gamma_u(\mathbf{b}, \mathbf{d})\eta_u}. \end{aligned} \tag{A2}$$

To derive the relations between the parameters λ and η, let us consider the probability of observing a particular response vector $\mathbf{X}_{obs}$. On the one hand, it is given in Equation (5). On the other hand, it can be presented as follows:

$$\begin{aligned} p(\mathbf{X}_{obs} = \mathbf{x}) &= \sum_{t=0}^{m} p(\mathbf{X}_{obs} = \mathbf{x}, X_{+mis} = t) \\ &= \sum_{t=0}^{m} \frac{\prod\limits_{i=1}^{n} b_i^{x_i} \gamma_t(\mathbf{d})\eta_{x_+ + t}}{\sum\limits_{u=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d})\eta_u} = \frac{\prod\limits_{i=1}^{n} b_i^{x_i} \sum\limits_{t=0}^{m} \gamma_t(\mathbf{d})\eta_{+ + t}}{\sum\limits_{s=0}^{n} \gamma_s(\mathbf{b}) \left(\sum\limits_{t=0}^{m} \gamma_t(\mathbf{d})\eta_{s+t} \right)}. \end{aligned} \tag{A3}$$

Hence,

$$\lambda_s = \sum_{t=0}^{m} \gamma_t(\mathbf{d})\eta_{t+s}, \forall s \in [0, n]. \tag{A4}$$

APPENDIX B

The probability of answering the new item correctly is:

$$\begin{aligned} \Pr(X_{mis} = 1) &= \sum_{\mathbf{x}} \Pr(X_mis = 1, \mathbf{X}_{obs} = \mathbf{x}) \\ &= \frac{d\sum_{t=0}^{n} \gamma_t(\mathbf{b})\eta_{t+1}}{d\sum_{t=0}^{n} \gamma_t(\mathbf{b})\eta_{t+1} + \sum_{t=0}^{n} \gamma_t(\mathbf{b})\eta_t} \end{aligned} \tag{A5}$$

Using the general solution of the system of equations in Equation (6), the two sums in this expression can be re-written as:

$$\begin{aligned} d\sum_{t=0}^{n} \gamma_t(\mathbf{b})\eta_{t+1} &= d\sum_{t=0}^{n} \left(\gamma_t(\mathbf{b}) \sum_{s=0}^{t} \frac{(-1)^{t-s}\lambda_s}{d^{t+1-s}} + (-1)^t \frac{k}{d^{t+1}} \right) \\ &= \sum_{t=0}^{n} \gamma_t(\mathbf{b}) \sum_{s=0}^{t} \frac{(-1)^{t-s}\lambda_s}{d^{t-s}} + k\sum_{t=0}^{n} \gamma_t(\mathbf{b}) \frac{(-1)^t}{d^t} \\ &= \sum_{t=0}^{n} \gamma_t(\mathbf{b}) \sum_{s=0}^{t-1} \frac{(-1)^{t-s}\lambda_s}{d^{t-s}} + \sum_{t=0}^{n} \gamma_t(\mathbf{b})\lambda_t + k\sum_{t=0}^{n} \gamma_t(\mathbf{b}) \frac{(-1)^t}{d^t} \end{aligned} \tag{A6}$$

and

$$\begin{aligned} \sum_{t=0}^{n} \gamma_t(\mathbf{b})\eta_t &= \sum_{t=0}^{n} \left(\gamma_t(\mathbf{b}) \sum_{s=0}^{t-1} \frac{(-1)^{t-1-s}\lambda_s}{d^{t-s}} + (-1)^{t-1} \frac{k}{d^t} \right) \\ &= -\sum_{t=0}^{n} \gamma_t(\mathbf{b}) \sum_{s=0}^{t-1} \frac{(-1)^{t-s}\lambda_s}{d^{t-s}} - k\sum_{t=0}^{n} \gamma_t(\mathbf{b}) \frac{(-1)^t}{d^t}. \end{aligned} \tag{A7}$$

Hence,

$$\begin{aligned} \Pr(X_{mis} = 1) &= \frac{\begin{array}{c} \sum_{t=0}^{n} \gamma_t(\mathbf{b}) \sum_{s=0}^{t-1} \frac{(-1)^{t-s}\lambda_s}{d^{t-s}} \\ + \sum_{t=0}^{n} \gamma_t(\mathbf{b})\lambda_t + k\sum_{t=0}^{n} \gamma_t(\mathbf{b}) \frac{(-1)^t}{d^t} \end{array}}{\sum_{t=0}^{n} \gamma_t(\mathbf{b})\lambda_t} = \\ &= 1 - \frac{k\sum_{t=0}^{n} \frac{(-1)^{t-1}\gamma_t(\mathbf{b})}{d^t}}{\sum_{t=0}^{n} \gamma_t(\mathbf{b})\lambda_t} + \frac{\sum_{t=1}^{n} \sum_{s=0}^{t-1} \frac{(-1)^{t-s}\gamma_t(\mathbf{b})\lambda_t}{d^{t-s}}}{\sum_{t=0}^{n} \gamma_t(\mathbf{b})\lambda_t}. \end{aligned} \tag{A8}$$

First, we will consider the constraints on k, following from the parameters $\boldsymbol{\eta}$ being positive:

$$\begin{cases} k = \eta_0 > 0, \\ \sum_{t=0}^{s-1} \frac{(-1)^{s-t+1}\lambda_t}{d^{s-t}} + (-1)^s \frac{k}{d^s} > 0, \forall s \in [1 : (n+1)]. \end{cases} \tag{A9}$$

For even indices $s = 2u$, $u = 1, 2, \ldots, \lfloor \frac{n+1}{2} \rfloor$, we have:

$$\frac{k}{d^{2u+1}} > \sum_{t=0}^{2u} \frac{(-1)^t\lambda_t}{d^{2u+1-t}} \Leftrightarrow k > \sum_{t=0}^{2u} (-1)^t\lambda_t d^t. \tag{A10}$$

For odd indices $s = 2u + 1$, $u = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$, we have:

$$\frac{k}{d^{2u}} < \sum_{t=0}^{2u-1} \frac{(-1)^t\lambda_t}{d^{2u-t}} \Leftrightarrow k < \sum_{t=0}^{2u-1} (-1)^t\lambda_t d^t. \tag{A11}$$

Second, we consider the monotonicity constraints (12): $\eta_{s+1}\eta_{s-1} > \eta_s^2, \forall s \in [1 : n]$. Using the the general solution of the system of equations in Equation (10), we have:

$$\begin{aligned} &\left(\sum_{t=0}^{s} \frac{(-1)^{s-t}\lambda_t}{d^{s+1-t}} + (-1)^{s+1} \frac{k}{d^{s+1}} \right) \\ &\qquad \left(\sum_{t=0}^{s-2} \frac{(-1)^{s-t-2}\lambda_t}{d^{s-1-t}} + (-1)^{s-1} \frac{k}{d^{s-1}} \right) \\ &\qquad > \left(\sum_{t=0}^{s-1} \frac{(-1)^{s-t-1}\lambda_t}{d^{s-t}} + (-1)^s \frac{k}{d^s} \right)^2. \end{aligned} \tag{A12}$$

If we multiply both sides by d^{2s} and denote $S = \sum_{t=0}^{s-2}(-1)^{s-t}\lambda_t d^t$, then we get

$$\left(S + \lambda_s d^s - \lambda_{s-1}d^{s-1} - (-1)^s k\right)\left(S - (-1)^s k\right) > \left(-S + \lambda_{s-1}d^{s-1} + (-1)^s k\right)^2. \quad \text{(A13)}$$

When multiplying the elements on the left side and taking a square on the right side, most of the element on the both sides are the same, hence they cancel out, and the remaining inequality is:

$$(S - (-1)^s k)\lambda_s d^s > -(S - (-1)^s k)\lambda_{s-1}d^{s-1} + (\lambda_{s-1}d^{s-1})^2 \Leftrightarrow$$
$$S - (-1)^s k > \frac{\lambda_{s-1}^2 d^{s-1}}{\lambda_s d + \lambda_{s-1}}. \quad \text{(A14)}$$

For even indices $s = 2u$, $u = 1, 2, \dots, \lfloor \frac{n}{2} \rfloor$, we have:

$$k < \sum_{t=0}^{2u-2}(-1)^t \lambda_t d^t - \frac{\lambda_{2u-1}^2 d^{2u-1}}{\lambda_{2u}d + \lambda_{2u-1}}. \quad \text{(A15)}$$

For odd indices $s = 2u + 1$, $u = 0, 1, \dots, \lfloor \frac{n-1}{2} \rfloor$, we have:

$$k > \frac{\lambda_{2u}^2 d^{2u}}{\lambda_{2u+1}d + \lambda_{2u}} + \sum_{t=0}^{2u-1}(-1)^t \lambda_t d^t. \quad \text{(A16)}$$

APPENDIX C

For a simulation approach (such as a Gibbs Sampler) to be applicable, we have to show that the solutions of the system of Equation (10) and inequalities (16) constitute a convex and bounded set, which ensures that the sampler can easily cover the full subspace of possible values of the non-identified parameters. All coefficients in the system of equations are positive, so are the parameters $\boldsymbol{\lambda}$, and therefore each of the parameters η_s is bounded:

$$0 < \eta_s < \min_{t=\max(0,\, s-m)}^{\min(s,n)} \frac{\lambda_t}{\gamma_{s-t}(\mathbf{b})}, \forall s \in [0 : (n+m)]. \quad \text{(A17)}$$

For every $s \in [1 : (n + m - 1)]$ the solutions of the following set of inequalities:

$$\begin{cases} \frac{\eta_{s+1}}{\eta_s} \geq \frac{\eta_s}{\eta_{s-1}} \\ \eta_{s-1} > 0 \\ \eta_s > 0 \\ \eta_{s+1} > 0 \end{cases} \quad \text{(A18)}$$

form a convex set. The interaction of convex sets from each s is itself a convex set. The intersection of the set formed by solutions of all inequalities and the set formed by the system of linear equations (which is always a convex set) is also a convex set. Therefore, all possible values of $\boldsymbol{\eta}$ constitute a convex set, and for each individual parameter there is only one range of possible values. Although the parameters are not identified, it is still possible to sample from their joint posterior distribution. The data augmented Gibbs Sampler for test equating with the ERM which is an extension of the algorithm of Maris et al. (2015) was developed for this. The details of our algorithm can be found in the Appendix E.

APPENDIX D

Non-Equivalent Group Equating Designs

The most simple non-equivalent group design one is the anchor-item design, in which both the reference and the new tests include a common set of items. The second design is a post-equating design, in which the link between the two tests is established through the data collected in the so called linking groups, answering some items from the reference test together with some items from the new test. The third design is a variation of the post-equating design, in which persons from some of the linking groups answer the items from the reference test and some other items from the item bank, while persons from the other linking groups answer the items from the new test and the same items from the item bank. The items that do not belong to either the reference test or the new test might also be taken by students from some historic population in one of the previous years. The simplest forms of these designs are visualized in **Figure A1**.

Let us by $\mathbf{Y}$ denote the $M \times m$ data matrix with responses of a sample of persons from the new population to the new test, by $\boldsymbol{\kappa}$ denote the $m + 1$ identified population parameters of the new population, by $\{r\}$ the set of items in the reference test and by $\{c\}$ the set of items in the new test. If the design includes linking groups, then $\mathbf{Z}$ is the data coming from the equating groups with $K^{(g)}$ and $k^{(g)}$ being the number of persons in the g-th linking group and the number of items answered by them; $\{e^{(g)}\}$ denotes the set of items answered by the g-th linking group and $\boldsymbol{\tau}^{(g)}$ are the population parameters of this linking group. Then the density of the observed data is:

$$f(\mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}) = \frac{\prod_i b_i^{u_i} \prod_{s=0}^{n}\left(\sum_{t=0}^{|c/r|}\gamma_t(\mathbf{b}_{c/r})\eta_{t+s}\right)^{N_s} \prod_{s=0}^{m}\kappa_s^{M_s}}{\left(\sum_{t=0}^{|r\cup c|}\gamma_t(\mathbf{b}_{r\cup c})\eta_t\right)^{N}\left(\sum_{t=0}^{m}\gamma_t(\mathbf{b}_c)\kappa_t\right)^{M}}$$
$$\prod_g \frac{\prod_{s=0}^{k^{(g)}}\left(\tau_s^{(g)}\right)^{K_s^{(g)}}}{\left(\sum_{t=0}^{k^{(g)}}\gamma_t\left(\mathbf{b}_{e^{(g)}}\right)\tau_t^{(g)}\right)^{K^{(g)}}}, \quad \text{(A19)}$$

where u_i is the total number of correct responses to item i by all students which answered this item, N_s, M_s, and $K_s^{(g)}$ are the number of persons from the reference population, new population and the g-th linking group, respectively, that gave exactly s correct responses to the items in the corresponding tests.

The score distribution of the reference population on the new test depends on the population parameters $\boldsymbol{\eta}$ and the item

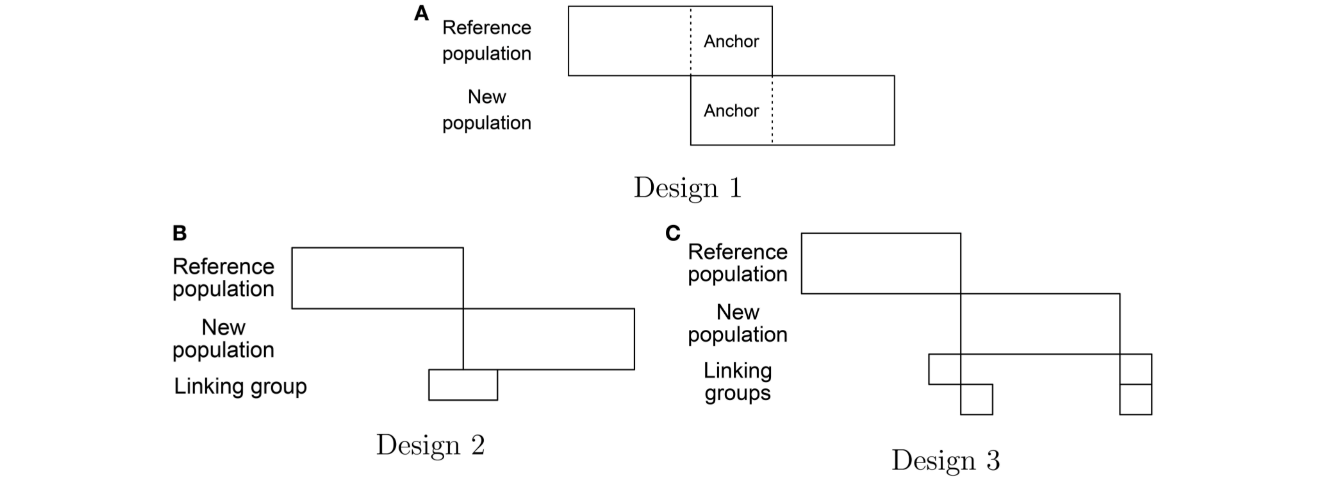

FIGURE A1 | Three non-equivalent group equating designs. (A) Design 1, **(B)** Design 2, **(C)** Design 3.

parameters $\mathbf{b}$:

$$p(X_{+mis} \leq T) = \sum_{t=0}^{T} \frac{\gamma_t(\mathbf{b}_{c/r})(\sum_{s=0}^{n} \gamma_s(\mathbf{b}_{r/c})\eta_{s+y})}{\sum_{u=0}^{|r \cup c|} \gamma_u(\mathbf{b})\eta_u} \tag{A20}$$

To make inferences about this distribution we obtain samples from the posterior distribution $p(\boldsymbol{\eta}, \mathbf{b} \mid \ldots)$. This is done using a data augmented Gibbs sampler.

APPENDIX E

We describe here how the samples from the joint posterior distribution

$$p(\boldsymbol{\eta}, \mathbf{b} \mid \ldots) \tag{A21}$$

can be obtained using a Markov chain Monte Carlo algorithm. We describe it for the post-equating non-equivalent groups design (see **Figure A1B**) with G linking groups. The density of the data given this equating design is given in Equation (A19) in Appendix D. The algorithm can be easily altered for the different kinds of non-equivalent group designs.

Data Augmented Gibbs Sampler

For computational convenience, instead of parameters $\boldsymbol{\eta}$, we use a different parametrization with the ratios of the consecutive parameters $\frac{\eta_{s+1}}{\eta_s}$. To place the parameters on the scale common in IRT we consider logarithms of these ratios:

$$p_s = \ln\left(\frac{\eta_{s+1}}{\eta_s}\right), \forall s \in [0 : n+m-1]. \tag{A22}$$

We use a prior which in addition to the monotonicity constraint (5) has a lower and an upper bound for the parameters:

$$p(\mathbf{p}) \propto \prod_{s=0}^{n+m-1} \mathcal{I}_{[p_{s-1}, p_{s+1}]}(p_s), \tag{A23}$$

where $p_{-1} = -100$ and $p_{n+m} = 100$. This is a reasonable constraint, since it follows from the Dutch identity (Holland, 1990) that

$$p_s = \ln\left(\mathcal{E}(\exp(\Theta) \mid X_{+obs} + X_{+mis} = s)\right). \tag{A24}$$

A priori, item and population parameters are independent. For item parameters we choose a uniform prior for difficulty parameters $-\ln(b_i)$, which is $p(b_i) \propto \frac{1}{b_i}$.

After the re-parametrization, the density of the observed data is:

$$f(\mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(G)}) = \prod_{g=1}^{G} \frac{\prod_{i \in \{r \cap e_g\}} b_i^{x_{+i}+z_{+i}^{(g)}} \prod_{i \in \{c \cap e_g\}} b_i^{y_{+i}+z_{+i}^{(g)}} \prod_{s=1}^{k^{(g)}-1} \exp(r_s^{(g)})^{\sum_{j>s} K_j^{(g)}}}{(1+\sum_{t=1}^{k^{(g)}} \gamma_t(\mathbf{b}_{e_g}) \prod_{j<t} \exp(r_j^{(g)}))^{K^{(g)}}} \times$$

$$\frac{\prod_{i \in \{r/e\}} b_i^{x_{+i}} \prod_{i \in \{c/e\}} b_i^{y_{+i}} \prod_{s=0}^{n} (1+\sum_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod_{j<t} \exp(p_j))^{N_s} \prod_{s=0}^{m-1} \exp(q_s)^{\sum_{j>s} M_j}}{(1+\sum_{t=1}^{n+m} \gamma_t(\mathbf{b}) \prod_{j<t} \exp(p_j))^N (1+\sum_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod_{j<t} \exp(q_j))^M}, \tag{A25}$$

where $\mathbf{p}, \mathbf{q}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(G)}$ are the population parameters of the reference population, the new population and G linking groups, respectively.

Although, we are interested only in the parameters $\mathbf{b}_c$ and $\mathbf{p}$, the other parameters $(\mathbf{b}_r, \mathbf{q}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(G)})$ are also sampled as nuisance parameters. Moreover, to make the full conditional posterior distribution of p_s tractable, at every iteration we will sample augmented data $\mathbf{x}^*$: responses of persons from the reference group to the items of the new test (Tanner and Wong, 1987; Zeger and Karim, 1991). This amounts to sampling from the joint posterior:

$$p(\mathbf{p}, \mathbf{q}, \mathbf{r}^{(1)}, \dots, \mathbf{r}^{(G)}, \mathbf{b}, \mathbf{x}^* \mid \mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(G)}). \qquad \text{(A26)}$$

A Gibbs sampler is used, i.e., all parameters are subsequently sampled from their full conditional distributions given the new values of all other parameters (Geman and Geman, 1984; Casella and George, 1992). After starting from the initial values (1 for all item parameters, and the population parameters equally distanced from –3 to 3), the algorithm goes through the following steps:

Step 1. Sample the augmented data $\mathbf{x}^*$.

For every person $j \in [1:N]$, sample a vector of responses $\mathbf{x}^*_j$ from its full conditional posterior $p(\mathbf{x}^*_j \mid \dots)$, which is factored in the following way:

$$\begin{aligned} p(\mathbf{x}^*_j \mid \dots) &= p(x^*_{j+} \mid x_{j+}, \mathbf{b}_c, \mathbf{p}) p(\mathbf{x}^*_j \mid x^*_{j+}, \mathbf{b}_c) \\ &= p(x^*_{j+} \mid x_{j+}, \mathbf{b}_c, \mathbf{p}) \\ &\quad \times p(x^*_{j,1} \mid x^*_{j+}, \mathbf{b}_c) p(x^*_{j,2} \mid x_{j,1}, x^*_{j+}, \mathbf{b}_c) \dots \\ &\quad p(x^*_{j,m} \mid x^*_{j,1}, \dots, x^*_{j,m-1}, x^*_{j+}, \mathbf{b}_c), \end{aligned} \qquad \text{(A27)}$$

where x_{j+} is the sumscore of person j. First, sample x^*_{j+} from the categorical distribution with probabilities

$$\Pr(x^*_{j+} = s \mid x_{j+}, \mathbf{p}, \mathbf{b}_c) = \frac{\gamma_s(\mathbf{b}_c) \prod_{u<(x_{j+}+s)} \exp(p_u)}{1 + \sum_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod_{u<(x_{p+}+t)} \exp(p_u)}. \qquad \text{(A28)}$$

And then for every item $i \in [1:m]$ sample $x^*_{j,i}$ from a Bernoulli distribution with probability:

$$\Pr(x^*_{j,i} = 1 \mid x^*_{j+}, \mathbf{x}^*_{j,s<i}, \mathbf{b}_c) = \frac{b_i \gamma_{x^*_{j+} - \sum_{s=0}^{i-1} x^*_{j,s} - 1}(b_{i+1}, \dots, b_m)}{\gamma_{x^*_{j+} - \sum_{s=0}^{i-1} x^*_{j,s}}(b_i, b_{i+1} \dots, b_m)} \qquad \text{(A29)}$$

Step 2. Sample from the full conditional posterior of the distribution of the item parameters.

For every $i \in \{r/e\}$, sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{x_{+i}-1}}{(1+cb_i)^N}, \qquad \text{(A30)}$$

where $c = \frac{\sum_{t=1}^{n+m} \gamma_{s-1}(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}{\sum_{t=0}^{n+m-1} \gamma_s(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}$. This is an scaled beta-prime distribution, to sample from which first sample $y = \frac{cb_i}{1+cb_i}$ from $\mathcal{B}(x_{+i}, N - x_{+i})$, and then transform it: $b_i = \frac{1}{c}\frac{y}{1-y}$.

For every $g \in [1:G]$, for every $i \in \{r \cap e_g\}$, sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{x_{+i} + z^{(g)}_{+i} - 1}}{(1+c_1 b_i)^N (1+c_2 b_i)^{K^{(g)}}}, \qquad \text{(A31)}$$

where $c_1 = \frac{\sum_{t=1}^{n+m} \gamma_{s-1}(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}{\sum_{t=0}^{n+m-1} \gamma_s(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}$ and $c_2 = \frac{\sum_{t=1}^{k^g} \gamma_{s-1}(\mathbf{b}^{(i)}_{eg}) \prod_{j<t} \exp(r^{(g)}_j)}{\sum_{t=0}^{k^{(g)}-1} \gamma_s(\mathbf{b}^{(i)}_{eg}) \prod_{j<t} \exp(r^{(g)}_j)}$. Unlike the full conditional of the item parameters of the items taken by persons from only one population, this distribution is not easy to sample from directly. It is more convenient to sample from the distribution of $\beta_i = -\ln(b_i)$ using a Metropolis-Hasting algorithm (Metropolis et al., 1953). We use $\mathcal{N}(-\ln(b_i), \tau^2 = 0.01)$ as a proposal density with b_i being the current value of the parameter.

For every $i \in \{c/e\}$, sample b_i from its full conditional posterior analogously to sampling b_i, $i \in \{r \cap e_g\}$, because these items are not only taken by the new population, but responses to these items by the reference population are imputed.

For every $g \in [1:G]$, for every $i \in \{c \cap e_g\}$ sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{y_{+i} + z^{(g)}_{+i} + x^*_{+i} - 1}}{(1+c_1 b_i)^N (1+c_2 b_i)^{K^{(g)}} (1+c_3 b_i)^M}, \qquad \text{(A32)}$$

where $c_3 = \frac{\sum_{t=1}^{m} \gamma_{s-1}(\mathbf{b}^{(i)}_c) \prod_{j<t} \exp(q_j)}{\sum_{t=0}^{m-1} \gamma_s(\mathbf{b}^{(i)}_c) \prod_{j<t} \exp(q_j)}$. Use the same Metropolis-Hastings algorithm as for the items, taken by two populations. If the equating design specifies more than 3 populations taking some of the items, then the full conditional posteriors of those items can be extended accordingly.

Step 3. Sample the population parameters.

For every $s \in [0{:}(n{+}m{-}1)]$, sample p_s from its full conditional posterior:

$$p(p_s \mid \dots) \propto \frac{\exp(p_s)^{\sum_{j>s} N^*_s}}{(1+c\exp(p_s))^N} \mathcal{I}_{[p_{s-1}, p_{s+1}]}(p_s), \qquad \text{(A33)}$$

where $c = \frac{\sum\limits_{t=s+1}^{n} \gamma_t(\mathbf{b}) \prod_{j \neq s, j=0}^{t-1} \exp(p_j)}{1 + \sum\limits_{t=1}^{s} \gamma_t(\mathbf{b}) \prod\limits_{j=0}^{t-1} \exp(p_j)}$. To sample from this distribution, we first sample $y = \frac{c\exp(p_s)}{1+c\exp(p_s)}$ from the truncated beta distribution

$$f(y) \propto y^{\sum_{j>s} N^*_s - 1} (1-y)^{N - \sum_{j>s} N^*_s - 1} \mathcal{I}_{[a_1, a_2]}(y), \qquad \text{(A34)}$$

where $a_1 = \frac{c\exp(p_{s-1})}{1+c\exp(p_{s-1})}$ and $a_2 = \frac{c\exp(p_{s+1})}{1+c\exp(p_{s+1})}$, using rejection sampling with $U(a_1, a_2)$ as a proposal distribution, and then transform it: $p_s = \ln(\frac{1}{c}\frac{y}{1-y})$.

For every $s \in [0:(m-1)]$, sample q_s from its full conditional posterior analogously to sampling p_s. For every $g \in [1:G]$, for every $s \in [0:k^{(g)}-1]$, sample $r^{(g)}_s$ form its full conditional posterior analogously to sampling p_s.

At every iteration of the Gibbs sampler, we compute the expected score distribution for the reference population on the new exam $Pr(X_{+mis} \leq T)$.