UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXXI

Advances in test equating: comparing IRT and Kernel methods and a new likelihood approach to equate multiple forms

Coordinatore del corso: Prof. Nicola Sartori

Supervisore: Prof.ssa Michela Battauz (Università Degli Studi di Udine)

Co-supervisore: Prof.ssa Marie Wiberg (Umeå Universitet)

Dottorando: Waldir Leôncio Netto

November 30, 2018

Abstract

Test equating is a statistical procedure to ensure that scores from different test forms are comparable and can be used interchangeably (González and Wiberg, 2017). There are several methodologies available to perform equating, some of which are based on the Classical Test Theory (CTT) framework and others are based on the Item Response Theory (IRT) framework.

After a short overview of latent trait models and test equating on Chapter 1, Chapter 2 of this thesis proposes a procedure to compare equating transformations originated from different frameworks. As example, we have compared Item Response Theory Observed-Score Equating (IRTOSE), Kernel Equating (KE) and IRT Kernel Equating (IRTKE) under different scenarios. Our results suggest that IRT methods tend to provide better results than KE even when the data are not generated from IRT processes. KE can provide satisfactory results if a proper pre-smoothing solution can be found, while also being much faster than IRT methods. For daily applications, we recommend observing the sensibility of the results to the equating method, minding the importance of good model fit and meeting the assumptions of the framework.

Within the IRT framework, if the statistical modeling of the scores from each test form is performed independently, their respective parameters will be on different scales and thus incomparable. Equating solves this problem by transforming item parameters so they are all on the same scale. Popular IRT methods for equating pairs of test forms include the mean-sigma, mean-mean, Stocking–Lord and Haebara (Kolen and Brennan, 2014). For multiple forms, it might be necessary to employ more elaborate methods which take into account all the relationships between the forms.

Chapter 3 addresses this issue, as we propose a new statistical methodology that simultaneously equates a large number of test forms. Our proposal differentiates itself from the current state of the art by using the likelihood function of the true item parameters and the equating coefficients to perform the simultaneous estimation of all equating coefficients and by taking into account the heteroskedasticity of the item parameter

estimates as well as the correlations between those estimates on each test form. Such innovations give this new method the potential to yield equating coefficient estimates which are more efficient than what is currently available in the literature, albeit at a computational cost due to its increased complexity. This is indeed what has been observed in some of the simulations performed. Greater estimation efficiency is especially important in situations involving item parameters with extreme values.

Sommario

La procedura statistica di equating di un test viene utilizzata per garantire che i punteggi di diverse versioni di un test siano comparabili e possano essere usati in modo intercambiabile (González and Wiberg, 2017). Esistono diverse metodologie disponibili per eseguire l'equating, alcune delle quali sono basate sulla teoria classica dei test e altre sono basate sulla Item Response Theory (IRT).

Dopo una breve panoramica sui modelli per variabili latenti e sull'equating nel Capitolo 1, nel Capitolo 2 di questa tesi si propone una procedura per confrontare le conversioni originate da diversi metodi di equating. Ad esempio, abbiamo confrontato i metodi Item Response Theory Observed-Score Equating (IRTOSE), Kernel Equating (KE) e IRT Kernel Equating (IRTKE) in diversi scenari. I nostri risultati suggeriscono che i metodi IRT tendono a fornire risultati migliori rispetto a KE anche quando i dati non sono generati da processi IRT. KE può fornire risultati soddisfacenti se si può trovare un buon modello, pur essendo molto più veloce dei metodi IRT. Per le applicazioni quotidiane, raccomandiamo di osservare la sensibilità dei risultati al metodo di equating, prestando attenzione all'importanza della bontà di adattamento del modello e del soddisfacendo le ipotesi alla base del metodo.

All'interno dell'approccio IRT, se la modellazione statistica delle diverse versioni di un test viene eseguita in modo indipendente, i rispettivi parametri saranno su scale diverse e quindi incomparabili. Il processo di equating risolve questo problema trasformando i parametri degli item in modo che siano tutti sulla stessa scala. I metodi IRT più usati nel caso di due versioni di un test includono il mean-sigma, mean-mean, Stocking–Lord e Haebara (Kolen and Brennan, 2014). Per più di due versioni di un test, potrebbe essere necessario utilizzare metodi più elaborati che tengano conto di tutte le relazioni tra di esse.

Il Capitolo 3 affronta questo problema, in quanto proponiamo una nuova metodologia statistica che esegue l'equating simultaneamente su un gran numero di versioni di un test. La nostra proposta si differenzia dallo stato attuale dell'arte usando la funzione di

verosimiglianza dei veri parametri degli item e dei coefficienti di equating per eseguire la stima simultanea di tutti i coefficienti di equating, tenendo conto dell'eteroschedasticità delle stime dei parametri degli item e della correlazione tra di essi. Tali innovazioni danno a questo nuovo metodo il potenziale di fornire stime di coefficienti di equating più efficienti di quanto attualmente disponibile in letteratura, sebbene a un costo computazionale dovuto alla sua maggiore complessità. Questo è effettivamente ciò che è stato osservato in alcune delle simulazioni eseguite. Una maggiore efficienza di stima è particolarmente importante in situazioni che coinvolgono parametri degli item con valori estremi.

To my dad, who made sure to cross the Atlantic every year to visit, check up on and encourage me. To my mom, a paragon of unconditional love with whom I share the passion for scientific inquiry.

Acknowledgments

Being grateful in public is always a daunting task. If you forget to mention someone you (or they) think you should have, you come out as an ungrateful human being. Some people overcome this by keeping it short and simple and make sure to only mention the most important people in their whole lives, while others try to make it as general as possible so that not even their pets are left out. I have spent a lot of time thinking about which path to take and have decided to chronologically name some of the people whose support was crucial in making this thesis happen. I'll try to mention as many people as I can, but I also want to make it short so people actually want to read this. Yes, this is risky, but I haven't done anything since the start of my Ph.D. other than stepping out of my comfort zone.

Before coming back to Academia, I worked as a statistician at the Brazilian Attorney-General's Office (AGU), so that's where I'll start. There were two people there who were very important in making this dream of mine come true. The first one is Caio Vasconcelos, my last boss at AGU and one of the most visionary and hard-working people I have ever had the pleasure to meet. I was lucky enough to have been accepted as a Ph.D. student in Padua while I was working with him; as a prospective academic in his early 30s himself, he spared no expense in helping me get through all the infamous Brazilian bureaucracy necessary to make my professional transition as smooth as possible.

The second person at AGU I wanted to thank is Claudio Sousa, a Physics professor who part-timed as my AGU colleague and personal astrophysicist (sorry, Neil deGrasse Tyson). Seeing him passionately talk about science and academic life in general made me realize this is something I really wanted to do.

Being a Ph.D. student can be a daunting endeavor, especially when you've been away from Academia for no less than 10 years and jump straight from a B.S. to a Ph.D. program. I would therefore like to thank my first-year tutor, Bruno Scarpa, for helping me find my path and stay sane during that critical first year. I also have to thank Nicola Sartori and Ruggero Bellio for helping me find a thesis supervisor who was a great match to my academic interests.

Speaking of my supervisors, I would like to thank Michela Battauz and Marie Wiberg for all the support, mentoring and patience during these two research years. I felt like we worked great together and hope I can one day be as inspiring to a student as you are to me. I am finishing this program more excited about academic life than I was when I started, which is something hard to happen unless you are working with great people.

As far as institutions go, I cannot forget to mention all the people I have met at Unipd, Umeå University and CEMO (University of Oslo). I have a hard time thinking of one academic or administrative staff member I've met at those institutions who hasn't helped me solve a problem. I can only hope to have been a boon to your lives as well. And thanks to Fondazione Cassa di Risparmio di Padova e Rovigo for the generous scholarship, which sustained me during this whole period.

Lastly and most importantly, I have to thank my wife, Denise Reis, for all the support, inspiration and teachings during this period. We have been together since college, and she has never left Academia except for maybe a couple of years (where she was still working as a researcher), so perhaps she should have been the first one to have been mentioned in this chronology. But then again I wouldn't be able to imagine her freaking out while reading the previous page and failing to find her name somewhere over there.

After we graduated in 2005, I was tired of studying and just wanted to go out, get a job and make money while I figured out what and even *if* I ever wanted to come back to Academia for a Masters degree. Denise, on the other hand, already knew very well what she wanted back then, and kept on track. Ten years later, she's likely among the youngest and most well-qualified statisticians in our field, and I have nothing but respect and admiration for the authority she'd already achieved before her 30s.

I took a different professional path than my wife, and I think it worked very well, because when I decided to come back to Academia, I had more than my now-romanticized recollection of life as an undergrad student to fuel my desire: I was married to a Ph.D. student who I saw struggle and then succeed with her research. It is wonderful to see our professional paths converge again one decade later.

Moving continents and countries every six months for three years straight is no easy task, and it is even harder when your partner also has an active career, but somehow we made it work for both of us. To my beautiful wife, I say thank you for your flexibility, for your willingness to take so many leaps of faith with me, for supporting me through countless hard times and for making sure we never forgot to celebrate the good ones. I couldn't ask for a better life partner.

Ok, that's enough mushiness. Now everybody buckle up: it's time for oodles of science!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An overview of latent trait models

In his seminal work, Lord (1953) considered the perils of using the “raw score”—the addition of the scores across all items in a test—as an estimate of the underlying characteristic—also called the “latent trait” or “ability”—of an individual. The estimation of latent traits through the application of tests is at the heart of Classical Test Theory (CTT) and Item Response Theory (IRT), two popular statistical frameworks for addressing measurement problems in fields such as psychology and education. The basic difference between these two frameworks is how one focuses on total test scores and the other focuses on the responses to the individual items.

1.1.1 Classical Test Theory

In CTT, the observed score (X) of an individual in a test is composed of their true score (T) and a residual error score (E). Since only X is observed, the equation $X = T + E$ is unsolvable unless some assumptions are made. Thus, we assume that:

1. T and E are uncorrelated;
2. the average E in the population of examinees is zero;
3. the Es between two parallel tests—those that measure the same content and for which examinees have the same true score—are uncorrelated.

Under these assumptions, T is the expected value of an examinee's observed score across a large number of repeated (e.g. parallel) testings (Crocker and Algina, 2008).

Over the years, several measurement models have used CTT as a basis, either revising, expanding or weakening the basic assumptions laid above. Lord and Novick (1967) note that CTT models may be also referred to as "weak models" due to how easy it is for data to meet their assumptions.

1.1.2 Item Response Theory

Like CTT, IRT targets the estimation of a subject's latent trait by measuring their performance in a test. Unlike CTT, though, IRT focuses on the individual answers to the items that compose the test. There is great flexibility in the characteristics of the item responses—they can be discrete, continuous, dichotomous, polytomous, ordered, unordered—, and the latent trait under measurement can have just one or multiple dimensions. The most common cases, however, deal with the measurement of a single, unidimensional ability through the administration of dichotomous items. The relationship between those items and the ability is typically modeled by a one-, two- or three-parameter logistic model, the latter of which is described below.

Let $\theta_i \in \mathbb{R}$ be the latent trait under measurement on individual i and $X_{ij} = \{0, 1\}$ the score of i in item j. Moreover, let $a_j > 0$, $b_j \in \mathbb{R}$ and $c_j \in [0, 1)$ be three parameters associated with item j, respectively known as the item discrimination, difficulty and pseudo-guessing probability. The probability of $X_{ij} = 1$ given θ_i, a_j, b_j and c_j can be expressed as

$$\Pr(X_{ij} = 1|\theta_i; a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp\left[-a_j(\theta_i - b_j)\right]}, \tag{1.1}$$

which is also called the Item Characteristic Function (ICF). If we fix the item parameters, then the relation between $\Pr(X_{ij} = 1)$ and θ_i can be represented by a shape akin to the one on Fig. 1.1, which is called the Item Characteristic Curve (ICC). On that particular example, we have set $a_j = 1, b_j = 0$, and $c_j = 0.2$. This can be verified by checking the position of the inflection point of the curve (located at (0, 0.6), i.e. the point where $\theta = b_j = 0$ and $\Pr(S_i = 1)$ is in the middle of the 0.2–1.0 range), its inclination on that point (given by $a_j = 1$) and the limit of the curve as θ_i approaches $-\infty$ (given by $c_j = 0.2$).

The model above is called the three-parameter logistic model, or 3PL, due to the existence of three different item parameters and to how they model the relationship between the probability of scoring 1 on this item and an ability θ through a logistic curve. Alternative formulations of Eq. (1.1) consider the inclusion of a constant D that multiplies

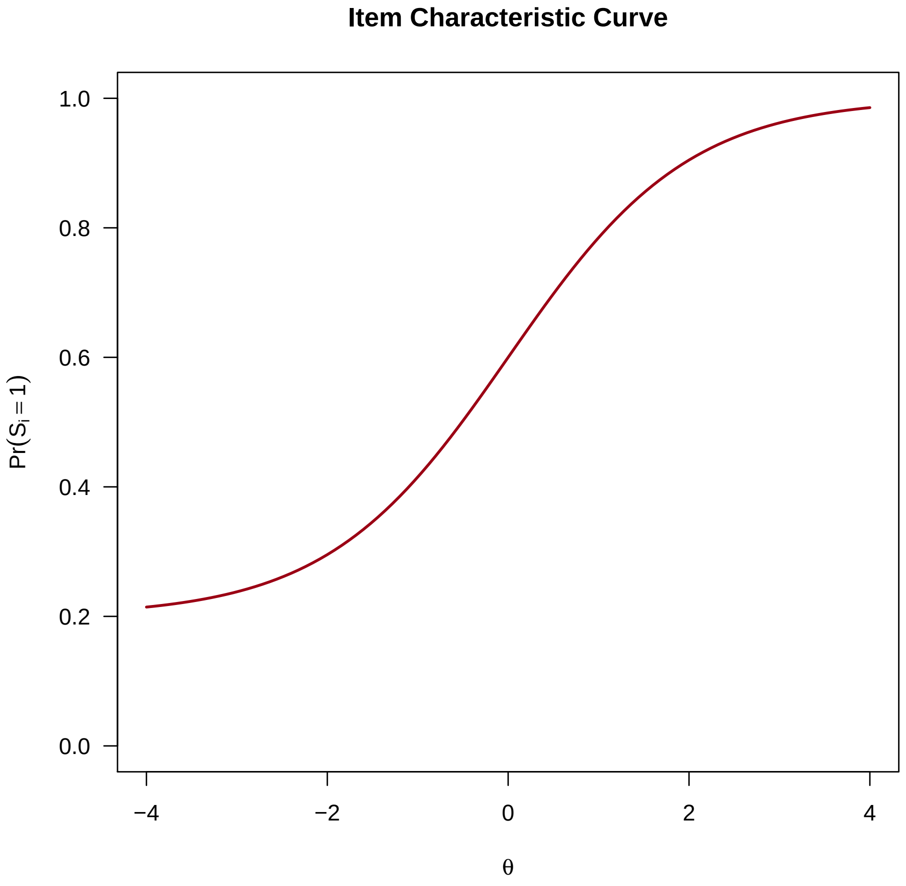


FIGURE 1.1: Item Characteristic Curve for $a = 1$, $b = 0$ and $c = 0.2$.

the parameter a_j. On this work, we consider $D = 1$ for simplicity, but it is also common to define $D = 1.7$ to make the ICC better approximate the Normal cumulative distribution function (CDF).

An alternative to ICFs derived from the logistic curve are those based on the standard Normal CDF, which results in models such as the 3PNO, or three-parameter Normal ogive model. The 3PNO may be similarly-shaped, but it is functionally different from the 3PL shown in Eq. (1.1). Due to their rarity, Normally-distributed IRT models will not be considered in this work.

The higher complexity of IRT models when compared with CTT models also make them potentially more informative. In particular, Hambleton and Jones (1993) note that since the item parameters can be interpreted on the same scale as the ability, one can easily determine the ability range where an item works best. In addition, they cite another feature of IRT models being the existence of item and test information functions, which can show the contribution of each particular item to the assessment of the latent trait and help measure the estimation errors of θ.

IRT may be focused on the item level, but the ICF for all the items in a test can be accumulated, resulting in the Test Characteristic Function (TCF). The TCF can be used to predict the total score of an examinee given their ability θ.

Another characteristic of IRT models is that they have stronger assumptions than CTT models. This means that CTT has the potential to be suitable to more cases than IRT. However, item and person parameters from an IRT model that *properly* fits the data will have sample invariance, meaning that the item parameters will not depend on the ability distribution of the examinees and the person parameters will not depend on the set of test items (Hambleton *et al.*, 1991, Ch. 2). This is not the case with CTT models, whose parameters can be sample-dependent. While advances on the original CTT framework saw the introduction of item *statistics* to represent their difficulty and discriminating power, those statistics are not sample-invariant like their *parameter* counterparts in IRT, meaning their usefulness decreases when the examinee sample differs in some unknown way from the population (Hambleton and Jones, 1993).

1.2 Test equating

Achievement tests are of interest to agents operating at different levels of a society. Individuals and institutions rely on these tests for certification and admission purposes; governments use them to implement and monitor the public policies they create.

In many cases, a test intent on measuring some latent trait of an individual—or a group of individuals—needs to be administered at different moments, so that the evolution of this measure over time can be observed. To have test results which are as authentic as possible, the test administrator may be interested in keeping the contents of the test in secrecy. One way to achieve this is by keeping access to the content on a must-know basis, which can be difficult and costly to implement. A more sensible approach consists of separating the population into groups—organized, for instance, by date and/or location of administration—and giving each group a slightly different test. Each of these versions of the test is called a “test form”, and even though this may solve the aforementioned security concerns, it raises a different issue: how can different versions of a test be compared? For instance, if one test form is—accidentally or otherwise—generally more difficult than another, then a person with a certain score on the harder test should be considered of higher ability than another person with the same score on the easier test.

This is where test equating comes in handy. According to Kolen and Brennan (2014), equating is a statistical procedure that adjusts the scores on test forms so that those scores can be used interchangeably.

A variety of designs can be used to collect data for equating, and choosing one in particular is a matter of satisfying both practical and statistical demands (Kolen and Brennan, 2014). One popular data collection design is called the Non-Equivalent group with Anchor Test, or NEAT (von Davier *et al.*, 2004, section 2.4). Its structure is summarized on Figure 1.2 and formally described below.

Let P and Q be two independent populations of examinees of which samples of size I_{P} and I_{Q} are taken. Let X and Y be two unique sets of test forms, respectively containing J_{X} and J_{Y} items, and A be a common test form containing J_{A} items. Test form $\mathrm{X}^+ = \{\mathrm{X}, \mathrm{A}\}$ will be administered to the sample of population P, and test form $\mathrm{Y}^+ = \{\mathrm{Y}, \mathrm{A}\}$ will be administered to the sample of population Q. This notation will come in handy in Ch. 2, when the NEAT designed is mentioned again.

With regard to the number of items on each form, let J be the collective item pool, i.e., $J = \{J_{\mathrm{X}}, J_{\mathrm{Y}}, J_{\mathrm{A}}\}$. Notice how, since X, Y and A have no items in common, $J = J_{\mathrm{X}} + J_{\mathrm{Y}} + J_{A}$. In other words, $J = \sum_t J_t$ for non-communicating test forms $t = \{\mathrm{X}, \mathrm{Y}, \mathrm{A}\}$. From set theory, we also have $J = (J_{\mathrm{X}^+} + J_{\mathrm{Y}^+}) - J_{\mathrm{A}}$, where J_{X^+} and J_{Y^+} are the number of items of test forms X^+ and Y^+, respectively. The latter notation may seem cumbersome, but it is useful to keep in mind for situations where anchor items are internal, which is the case for the datasets in Ch. 3. In these cases, the administered test forms are identified as X^+ and Y^+, or simply as X and Y, with no explicit reference to anchor test form A. Consequently, this means that $J < \sum_t J_t$.

To equate tests X^+ and Y^+, we need to know not only the data collection design, but also the equating method used (von Davier *et al.*, 2004). In the NEAT data collection design, there are different ways of using the information provided by the anchor test forms to equate two tests (von Davier and Chen, 2013). Kolen and Brennan (2014) divide them into linear methods and equipercentile methods, with notable examples of the former being the Tucker method (Gulliksen, 1950), the Levine observed score method (Levine, 1955; Kolen and Brennan, 2014) and the Levine true score method (Levine, 1955); among the equipercentile methods, the frequency estimation method (Angoff, 1971; Braun and Holland, 1982) is highlighted.

In this work, we only consider the method of frequency estimation for equipercentile equating. Reasons for this include the fact that the linear methods cited above either require distributional assumptions that prevent their universal application or they are

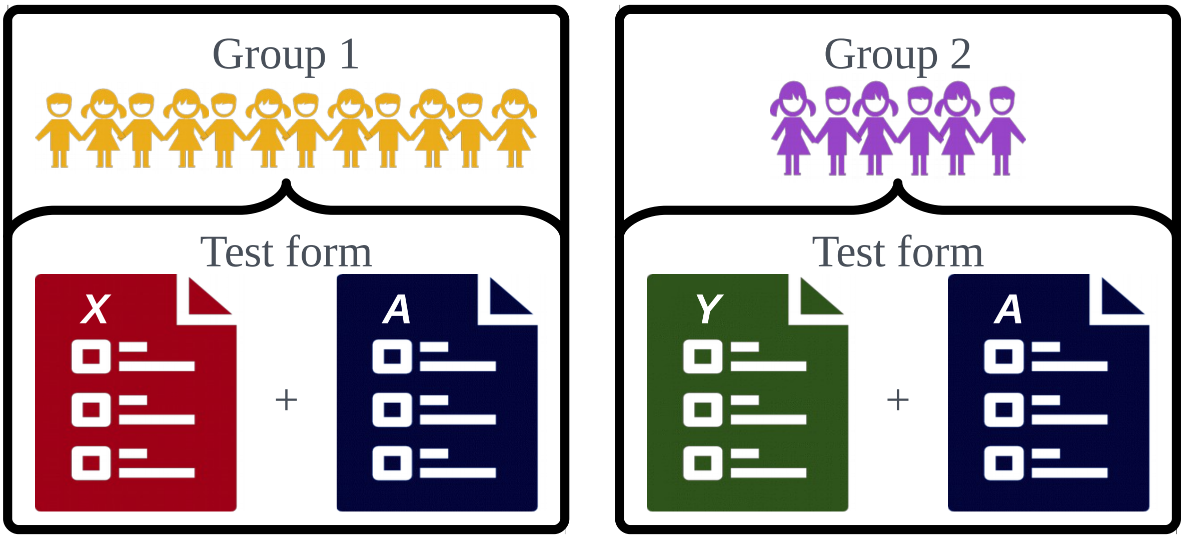


FIGURE 1.2: Structure of the Non-Equivalent groups with Anchor Test design. Group 1 and Group 2 come from populations P and Q, respectively, and may not have the same size. Test forms X, Y and A have no items in common, but the composed test forms $\mathrm{X}^{+} = \{\mathrm{X}, \mathrm{A}\}$ and $\mathrm{Y}^{+} = \{\mathrm{Y}, \mathrm{A}\}$ do.

tied to CTT, impairing their direct application to IRT-modeled data. In any case, all these methods can be used as part of larger equating structures which cover the entire spectrum of the equating procedure, from treating the observed test data to obtaining the equated scores.

In this thesis, we shall call the aforementioned equating structures “equating frameworks” or simply “frameworks”, which belong to but should not be confused with the two *latent-trait* frameworks of CTT and IRT. Two of these equating frameworks, namely Kernel Equating and IRT equating, are introduced in the following subsections. It should be noted, finally, that these equating frameworks are not inflexible: they contain parameters which can be tweaked, and each of those tweaks creates what we shall call a different “method” of that particular framework.

1.2.1 Kernel Equating

Kernel Equating (KE, von Davier *et al.*, 2004) is a framework comprising of several methods which have five steps in common: pre-smoothing, estimation of score probabilities, continuization of discrete distributions, equating, and calculation of evaluation measures.

On the first step, a statistical model—usually of polynomial log-linear form (Hanson, 1996)—is fitted to the observed data (raw scores per individual). Together with a

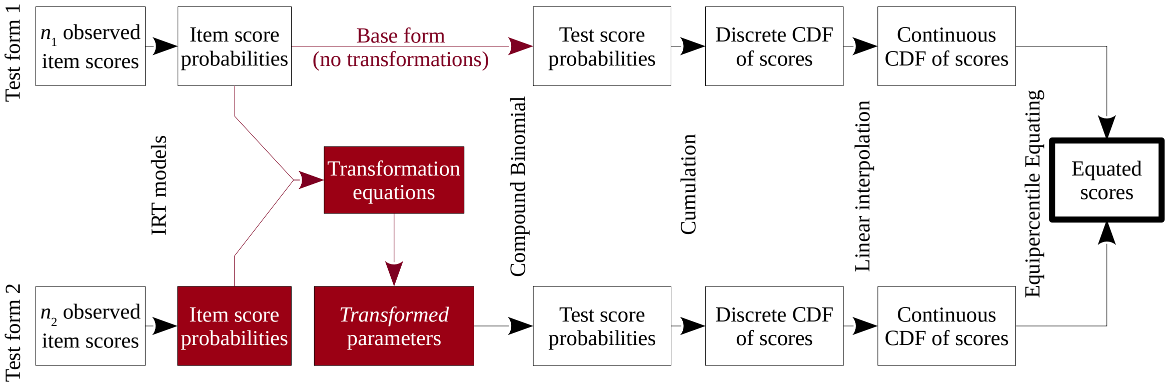


FIGURE 1.3: Workflow of the IRT observed-score equating process. Each row corresponds to a test form, and the colored part highlights the most critical part of the equating process, which is the transformation of the item parameter estimates.

design function, the pre-smoothed data is then used to estimate score probabilities. On the third step, a kernel function is used to transform those score estimations from a discrete to a continuous scale. Once the cumulated score probability distributions have been made continuous, the test forms can be equated by finding the scores that share a percentile, a straightforward process called "equipercentile equating" (Braun and Holland, 1982). Finally, accuracy measures such as the standard error of equating (SEE) can be calculated.

Since KE uses only the information regarding the total test scores, it can be seen as a CTT-based equating framework. This means it inherits all the advantages and disadvantages of CTT models we have discussed previously. For more details on how KE works and how it compares to other methodologies, see Ch. 2.

1.2.2 IRT equating

Just like KE, IRT equating (IRTE, Lord, 1980) comprises a set of methods based on the IRT framework. One clear difference between these two frameworks is that, unlike KE methods and their free-form models to estimate *test* score probabilities, IRTE usually deploys models with a well-defined form—see Eq. (1.1) back on page 2—to estimate *item* score probabilities.

Nonetheless, modeling is just the first step of the whole equating process in IRTE and KE. For IRT and the particular case of separate calibration—which is further discussed on Ch. 3—, the equating workflow is illustrated in Fig. 1.3 and briefly described below.

After modeling the item responses, one test form must be chosen to be the base form. As long as the symmetry property of equating is satisfied, it does not matter which form is the base (Lord, 1980). The item parameters of this base form will not change, but those of the other forms will be transformed so that the person and item parameters of those forms are on the same scale of the base form. This is essentially what happens in the colored part of Fig. 1.3, with the arrows leading to the "transformed parameters" box highlighting that the item parameters on the equated form are transformed based on the parameters of the items it has in common with the base form. The transformation equations can be found on Kolen and Brennan (2014, Sec. 6.2.1) and in Sec. 3.1.1 of this thesis.

If a test is scored using estimated IRT abilities, there is no need to go further than the transformation of the item parameters and person abilities, as these transformations mark the end of the IRT equating process (Kolen and Brennan, 2014). However, to make a parallel with KE, let us consider that one is interested in converting abilities—which are unbounded numbers along the $\mathbb{R}$ line—to something comparable to the raw scores, which are often non-negative, rational numbers.

By definition, IRT models are applied at the item level. Hence, it is necessary to find a way to eventually aggregate those individual item probabilities so a probability distribution of the test scores can be determined. This is usually done using the compound binomial distribution (Birnbaum, 1968) or an iterative process derived from it (Lord and Wingersky, 1984).

At this point, we are still dealing with discrete score distributions, which are made continuous so methods like equipercentile equating can be optimally implemented. In IRTE, this continuization is usually done through simple linear interpolation. After that, equating two forms is basically a matter of comparing scores on equivalent percentiles on the CDF of each test form score. Algebraically, the equipercentile transformation is given by Eq. (2.1) on page 11. Chapters 2 and 3 readdress and further explore the process behind IRT equating.

1.3 Main contributions of the thesis

The wide array of equating frameworks and methods currently available has naturally created the necessity to compare and choose the best one for a certain situation. This process can be challenging, especially if the equating methods under scrutiny belong to different frameworks, but it is still a possible and worthwhile endeavor if proper assumptions are made (Wiberg and González, 2016). On the next chapter, we propose a procedure to compare equating transformations from different frameworks, particularly between methods from KE and IRTE—introduced above—, as well as IRTKE, a hybrid of the previous two. Our novel proposal expands on the work of Wiberg and González (2016), who developed a methodology to compare equating methods from a common framework. This technology has immediate application in situations where KE and IRTE methods are being considered, which by itself contributes to the discussion of whether CTT or IRT is the most proper framework for a particular scenario.

Another innovation we have pursued in this thesis involves the equating of a large number of test forms. This is a common problem in large-scale assessment tests, often administered by governments and large institutions to measure educational quality at country and worldwide levels, usually with the intent of monitoring and improving education-related public policies.

The final part of this thesis aims to contribute to the development of equating in this top-level, high-stakes scenario. In Ch. 3, we propose a new statistical methodology to simultaneously equate test forms in any scenario within the NEAT design, especially those involving a large number of test forms. Our proposal differentiates itself from the current state-of-the-art works of Haberman (2009); Battauz (2013, 2017a) by using the likelihood function of the true item parameters and the equating coefficients to perform the simultaneous estimation of all equating coefficients and by taking into account the heteroskedasticity of the item parameter estimates as well as the correlations between these estimates on each test form. These innovations should yield equating coefficient estimates which are more efficient than what is currently available in the literature. This is especially important in situations involving item parameters with extreme values.

Chapter 2

Evaluating equating transformations in IRT observed-score and kernel equating methods

2.1 Introduction

Equating methods are used to ensure that scores from different test forms are comparable and can be used interchangeably (Kolen and Brennan, 2014; González and Wiberg, 2017). To obtain comparability, an equating transformation is used to map scores from one test form onto the scale of the other test form. Let X and Y denote the scores from test forms X and Y, respectively. We are interested in transforming X to the scale of Y. The general transformation function for comparing two samples or distributions of random variables is defined as

$$\varphi(x) = F_Y^{-1}[F_X(x)], \tag{2.1}$$

an equation commonly referred to as the equipercentile transformation (Braun and Holland, 1982). Different equating methods have been developed depending on the data collection design and the assumptions placed on the data. Examples of equating methods are traditional equating methods (Kolen and Brennan, 2014), observed-score kernel equating methods (von Davier *et al.*, 2004), Item Response Theory (IRT) methods (Lord, 1980), local equating methods (van der Linden, 2011), as well as mixtures of them as for example local kernel IRT equating (Wiberg *et al.*, 2014).

Recently, Wiberg and González (2016) used a statistical approach to show how one can compare equating transformations within a particular framework. They illustrated their approach within the Kernel Equating (KE) framework and discussed how it could be done within IRT Observed-Score Equating (IRTOSE) and local equating. To propose how to evaluate an equating transformation *within* an equating framework was thus an important step although relatively straightforward.

A remaining problem pointed out by Wiberg and González (2016) was how to evaluate equating transformations *between* different equating frameworks. This chapter concentrates on this problem, aiming to propose how to evaluate equating transformations which come from different frameworks. In particular, this chapter will focus on how to evaluate equating transformations from IRTOSE (Lord, 1980), KE (von Davier *et al.*, 2004), and IRT Observed-Score Kernel Equating (IRTKE, von Davier, 2010; Wiberg *et al.*, 2014; Andersson and Wiberg, 2017). Both simulated and real data will be used to conduct this study.

There are a number of equating methods and the underlying equating estimator can be parametric, semiparametric or nonparametric (González and von Davier, 2013; González and Wiberg, 2017). To evaluate which equating estimators should be used in different situations, statistical tools are needed. A common practice has been to use *equating-specific measures* to evaluate an equating transformation. Thus, depending on which equating framework is used, different measures have been employed to evaluate the equating transformation. A common feature of equating-specific evaluation measures is that they target different parts of the equating process and thus aim to evaluate the equating based on different but specific aspects.

An example of an equating-specific measure in KE is the percent relative error (PRE), which compares the moments in the observed and equated score distributions (Jiang *et al.*, 2012; von Davier *et al.*, 2004). In traditional methods it has been common to use the difference that matters (DTM), which was originally defined as the difference between equated scores and scale scores that are larger than half of a reported score unit (Dorans and Feigenbaum, 1994).

Summary indices as described in Han *et al.* (1997) have also been used. The summary measures typically use one particular equating transformation as standard and compare other equating transformations against it. The idea is to measure discrepancies between equivalent scores for two different equating methods. Both Harris and Crouse (1993) and Kolen and Brennan (2014) summarize traditional equating evaluation criteria as well as describe implementations of them. It is important to point out that even though equating transformations and different evaluation criteria exist there is no single criterion which is overall preferable (Harris and Crouse, 1993; Kolen and Brennan, 2014; Wiberg and González, 2016).

To provide a fair comparison of equating transformations from different frameworks, the most important part is to set up the comparison so it does not favor any particular framework. This is challenging, as small decisions made at each step of the equating process could in the end favor one of the frameworks unintentionally. For example, a common practice for test constructors is to model test items with IRT models to examine the item characteristics. Obviously the fit to a specific IRT model can vary between the items and this may ultimately have an impact on the equating of the test. In KE, on the other hand, the pre-smoothing step typically involves log-linear modeling of the item score probabilities.

An interesting question is whether KE works better than IRTOSE if the data is not generated from an IRT model. Another interesting question is how to generate item responses in a simulation study if we do not want to assume a particular underlying IRT model as that might affect the equating results. A possible approach to handling these problems, which is used in this chapter, is to implement multiple data-generating methods to attempt making a comparison which is as fair as possible to the different frameworks.

The rest of this chapter is structured as follows. In the next section, the equating methods used are described. Then, statistical evaluation criteria are presented, with the chosen equating parameters being described. The fourth section exhibits the characteristics of the real and the simulated data used to implement the equating methods under study. The results are given in the fifth section, and the last section contains some concluding remarks.

2.2 Methodology

2.2.1 The NEAT equating design

To perform observed-score equating, two components must be known: the data collection design and the equating method used (von Davier *et al.*, 2004). Let P and Q be two independent populations of examinees of which samples of size I_{P} and I_{Q} are taken. Let X and Y be two unique sets of test forms, respectively containing J_{X} and J_{Y} items, and A be a common test form containing J_{A} items. Test form $\mathrm{X}^{+} = \{\mathrm{X}, \mathrm{A}\}$ will be administered to the sample of population P, and test form $\mathrm{Y}^{+} = \{\mathrm{Y}, \mathrm{A}\}$ will be administered to the sample of population Q. This data collecting design is called non-equivalent groups with anchor test (NEAT, von Davier *et al.*, 2004, section 2.4).

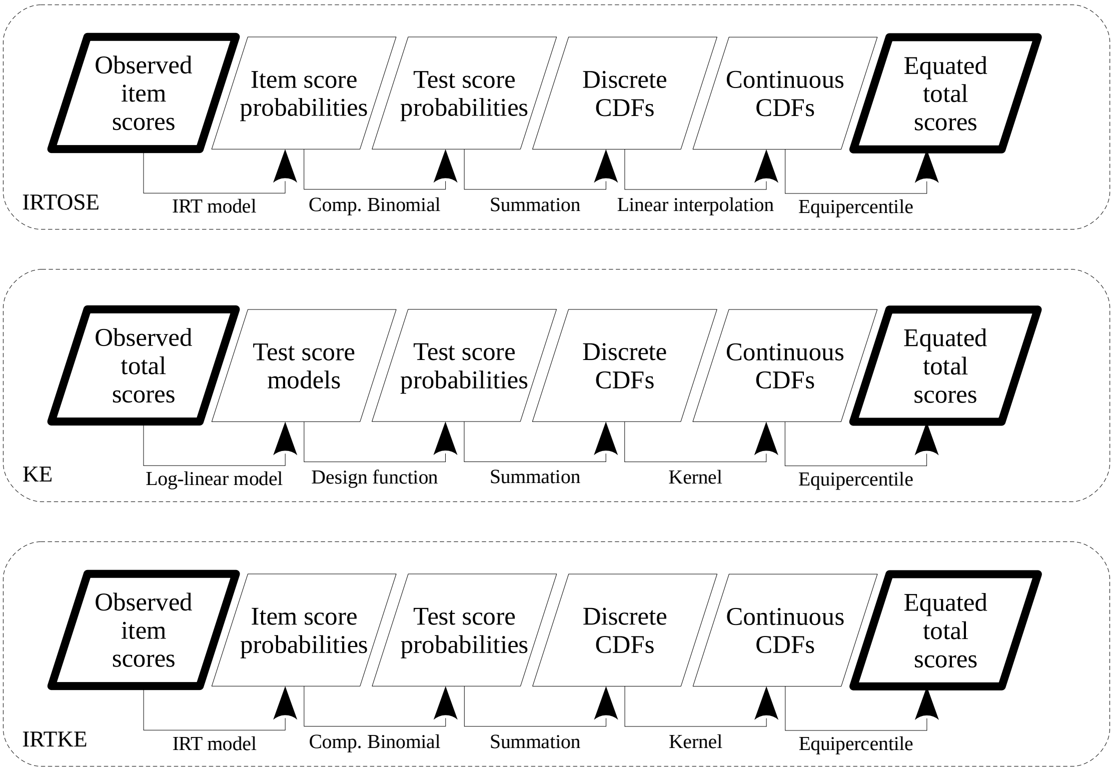


FIGURE 2.1: Simplified overview of three equating frameworks: IRTOSE, KE and IRTKE.

What follows is a brief description of the three equating frameworks studied in this chapter. To facilitate the methodological comparison between them, please refer to Fig. 2.1.

Differently from Fig. 1.3, here all the test forms to be equated are contained within one box for simplification. The workflow for all forms are identical in the steps shown and the difference between the base and non-base forms within a framework has been covered in Fig. 1.3 for IRTOSE; for KE and IRTKE, the workflow is the same for base and non-base forms.

Comparing the three frameworks in more general terms, what changes from one method to another are the input data needed, the statistical model applied to estimate the probability distribution of the test scores and the method for making such distribution continuous. After the continuous cumulative density function (CDF) of each test form is obtained, equipercentile equating is used to equate the observed scores.

2.2.2 IRT Observed-Score Equating (IRTOSE)

Let the observed data be composed of a matrix for each test form to be equated. Each one of those matrices is composed of I rows and J columns, where I is the number of examinees and J is the number of items on that same test form. Within those matrices we find dichotomous (correct/incorrect) answers to each one of those $I \times J$ combinations.

The application of IRTOSE begins with fitting an IRT model to the observed data. Let $X_{ij} = \{0, 1\}$ be the score of examinee i on item j. An IRT model calculates the probability of $X_{ij} = 1$ given the examinee's ability (θ_i) as well as some item parameters such as its discrimination (a_j) and difficulty (b_j). One of such models is the two-parameter logistic IRT model (2PL), which is defined as

$$\Pr(X_{ij} = 1 | a_j, b_j; \theta_i) = p_{ij} = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}. \tag{2.2}$$

Let X_i be the number-correct score (or simply "score") of examinee i, i.e.,

$$X_i = \sum_{j=1}^{J_\mathrm{X}} X_{ij} \tag{2.3}$$

and $X_i \in \{0, \ldots, J_\mathrm{X}\}$. A common way to calculate score probabilities is through a compound binomial model, here defined as

$$\Pr(X_i = x | \theta_i) = \sum_{\sum x_{ij} = x} \left[\prod_{j=1}^{J_\mathrm{X}} p_{ij}^{x_{ij}} (1 - p_{ij})^{1 - x_{ij}} \right]. \tag{2.4}$$

The compound binomial was derived by Birnbaum (1968); in practice, calculation is often performed through an iterative process described by Lord and Wingersky (1984), although other alternatives exist (González *et al.*, 2016).

The test score probabilities in Eq. (2.4) are still dependent on the abilities, so they must be marginalized over the sample to produce the probability distribution of the scores—$P(X_i = x)$—for a particular form. Since we are dealing with discrete score distributions, finding equivalent percentiles between two test forms will likely result in multiple solutions. Hence, these probability distributions are often made continuous. This is done by linear interpolation in IRTOSE. After continuization, equating two test forms is straightforward by applying Eq. (2.1) and thus finding the values on both forms that represent the same percentile.

IRTOSE is a flexible equating method which can be used with any data equating design, provided that the two test forms jointly fit an IRT model (Wiberg, 2016). On the flip side, its use of linear interpolation has a drawback: as pointed out by Andersson and Wiberg (2017), it does not provide an everywhere differentiable equating function, which is needed for the calculation of standard errors through response functions, as described by Ogasawara (2001).

2.2.3 Kernel Equating (KE)

KE is a framework comprised of the following five steps:

1. Pre-smoothing;
2. Estimation of score probabilities;
3. Continuization of discrete distributions;
4. Equating;
5. Calculating evaluation measures.

The goal of pre-smoothing is to fit a model to observed integer scores so that design functions can be better used to calculate the score probabilities. An especially useful family of pre-smoothing models are the log-linear models described in Rosenbaum and Thayer (1987) and Holland and Thayer (1987, 2000): they are well-behaved, relatively easy to estimate and flexible enough to fit the types of score distributions that arise in practice (von Davier *et al.*, 2004). Following Andersson *et al.* (2014), we will use a log-linear model fit through a Generalized Linear Model (GLM) for Poisson responses to model our data.

Once the observed distributions have been pre-smoothed, the scores of the two forms need to be linked. This can be done through chain equating (CE) or post-stratification equating (PSE). Choosing between the two methods is still largely an open research topic, but differences between their results tend to be negligible when the populations P and Q have similar distributions to the anchor test form A or when A correlates highly with both X and Y (von Davier *et al.*, 2004, section 11.8). In any case, the estimation of score probabilities is done by using a design function to transform the smoothed score distributions into marginal distributions for the populations.

In the third step, continuization of the discrete score distribution, KE uses a kernel—typically Gaussian, but also logistic, uniform or other—instead of linear interpolation. As with IRTOSE, once the cumulated score probability distributions have been made continuous, the two test forms can be equated by finding the scores located at the same percentile. Finally, accuracy measures such as the standard error of equating (SEE) can be calculated.

When compared to IRTOSE, KE offers the advantage of skipping the necessity to have the response matrix and to estimate parameters at the item level, not to mention the discussion about the most appropriate IRT model to use. Hence, KE has the potential to be less computationally intensive and more applicable than IRTOSE. On the other hand, it involves choosing (or not) a log-linear model to smooth the data, which arguably involves dealing with a greater range of possibilities than what IRT currently provides. Moreover, continuization requires selecting a kernel function as well as a smoothing parameter, which greatly increases the number of available methods. As for the statistical properties of KE and IRT, Meng (2012) observes that, under some conditions, KE seems more stable but not as accurate as IRT equating.

2.2.4 IRT Observed-Score Kernel Equating (IRTKE)

IRTKE, described by Andersson and Wiberg (2017), uses score probabilities derived from an IRT model as input for kernel continuization. It can be seen as a compromise between the typical IRTOSE and KE procedures.

Since this method uses the IRT models from Eqs. (2.2) and (2.4) on the pre-smoothing part of KE, it requires access to the item responses at the examinee level so that the model can be fit. When compared with CTT-based pre-smoothing models which only take into account the total test score of each individual, item-level models are more complex and consequently add computational overhead to the method.

As this framework uses a continuous and differentiable kernel instead of linear interpolation for continuization of the score distribution, it bypasses the issue of non-differentiability that can occur in IRTOSE methods. Andersson and Wiberg (2017) noticed that IRTKE works well for sample sizes as low as 1 000, as long as the 2PL model is used.

2.2.5 Choices for the simulation and real data study

Performing equating in one particular framework implies making several choices which result in one specific method within that framework. Picking only one method within each of the three frameworks under study is a decision that makes the number of comparisons manageable. One could argue that a fairer comparison between frameworks would require the evaluation of several methods for each framework, but our experience has been that the only case where changing the method gave wildly different final results was when a particular method generated blatantly unexpected equating results. This is usually a consequence of a model that clearly does not fit the data or simply fails to converge to a unique solution.

For IRTOSE and IRTKE, a 2PL model was fit to the item answers. When compared with alternative IRT models, the 2PL offers a good compromise between the simplicity of the 1PL and the flexibility of the 3PL. The flip side of simplicity is the potential of failure to capture important characteristics of the items; on the other hand, flexible models with many parameters may have problems with convergence, not to mention that the linking equations of the 2PL and the 3PL are identical. Since IRT is performed on each test form separately, the estimated item parameters and abilities are on incomparable scales that do not reflect the relationship between the test forms they model. The Stocking–Lord method (Stocking and Lord, 1983) was used to transform these item parameters.

To perform KE, several log-linear models were considered, ranging from simple functions containing only the scores of the main and anchor tests as covariates to complex ones containing several powers of the partial scores, the interactions between them, and dummy variables for low-frequency scores. The best model was then chosen by a stepwise method, which selected the model with the lowest Akaike Information Criterion (AIC). The linkage between the scores of the two forms was done through CE, and continuization was achieved with a Gaussian kernel. The same choices were made for IRTKE.

The performance of a particular equating framework can be affected not only by the method parameters set, but also by how the data behaves. Hence, we evaluated IRTOSE, KE and IRTKE on four different data-generating scenarios: a Swedish college admissions test, a Brazilian school assessment test, a simulated test generated from IRT parameters and another simulation with scores generated from Beta distributions.

All statistical procedures were performed in R (R Core Team, 2018), with `ltm` (Rizopoulos, 2006) being used to fit IRT models to the data and `glm` handling the log-linear models. IRTOSE was performed by `equateIRT` (Battauz, 2015); KE and IRTKE were done in `kequate` (Andersson *et al.*, 2013).

2.3 Evaluating equating transformations

The most common way to compare the performance of two equating transformations *within* a particular framework is through equating-specific evaluation measures (Wiberg and González, 2016). Two popular examples of those are the DTM, often used in traditional equating methods, and the PRE, which was specifically developed for KE but could be adapted to methods using linear interpolation (Jiang *et al.*, 2012). These measures could be adapted to compare equating transformations from different methods, but they are not examined further here.

In contrast to equating-specific measures, we could consider an equating transformation as a form of statistical estimator and calculate measures such as bias, standard error and mean square error (MSE). The advantage of this method is the familiarity of such measures, and their application to a between-framework scenario seems straightforward.

From Wiberg and González (2016), we respectively define the bias and MSE for an equated value $\varphi(x)$ of score x over R replications as

$$\mathrm{bias}[\hat{\varphi}(x)] = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{\varphi}^{(r)}(x) - \varphi(x)\right] \tag{2.5}$$

and

$$\mathrm{MSE}[\hat{\varphi}(x)] = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{\varphi}^{(r)}(x) - \varphi(x)\right]^2, \tag{2.6}$$

where $\hat{\varphi}^{(r)}(x)$ is the estimated equated score for the r-th replication.

The equations above make it clear that to calculate such measures we must have access to the true equating transformation $\varphi(x)$, which is not directly observable in real and even simulated data. There are, however, some ways to circumvent this limitation, one of which is to define one equating transformation as the true one and compare the others against it, something Wiberg and González (2016) did within KE. This chapter uses different approaches depending on whether we were dealing with real or simulated data. These procedures are summarized in the following subsections, and the complete code can be obtained upon request.

It is fair to assume that in certain situations a fast algorithm is preferable, even if it offers more bias and/or variance. Hence, runtimes of different equating frameworks are also confronted.

2.3.1 Choices for the real data

In order to calculate evaluation measures for the real data, we used the same approach employed by Lord (1977, p. 132), which basically consists of having test forms X and Y consist of the same items, while still having the computer handle them as being different. This procedure is summarized as follows:

1. The $I \times J_{\mathrm{X}}$ matrix containing the items and examinee answers for test form X was horizontally split into two matrices. The new matrices had half the number of examinees and the same number of items;
2. One of the resulting matrices was reassigned as test form Y;
3. Since X and Y are the same test form, equated scores should not change, i.e., $\varphi(x) = x$.

It should be noted that even if $\varphi(x)$ is expected to equal x, the observed differences will be different from zero not only due to bias, but also due to sample variability. Hence, for the real data, calling the measure $\hat{\varphi}(x) - \varphi(x)$ simply "bias" fails to recognize the effect of sampling error. For that reason, this error measure for the real data studies will be referred by the more comprehensive term "error". Moreover, since each real data scenario represents only one sample, the MSE would be simply the squared error calculated over one sample, which is improper and thus not calculated.

2.3.2 Choices for the simulated data

The advantage of working with computer simulations is that they give the user control over the parameters of the process that generates the data. The true equating scores are not explicitly defined by the data-generating parameters, but they can be obtained from them. Referring back to Fig. 2.1 on page 14, as long as the data-generating parameters allow the calculation of the expected CDFs of each test form, equipercentile equating can be performed.

For IRT-generated data, observations are generated by the true parameter values (a_j and b_j) and the true examinee skills (θ_i). Given these parameters, the score probability CDFs and the real score equating transformations $\varphi(x)$ were calculated as follows:

1. Eq. (2.2) was applied to calculate the probability of correctly answering an item, $\Pr(X_j = 1|a_j, b_j; \theta)$;
2. These probabilities were used in the compound binomial distribution to calculate the probability distribution of the test scores given the ability, $\Pr(X = x|\theta)$;
3. By integrating θ out of the probability above, one obtains the unconditional score probabilities $\Pr(X = x)$, which can be cumulated to form the (discrete) CDF for one form.
4. Once the CDFs for forms X and Y are calculated, the equivalent scores between them are found through equipercentile equating.

It is worth noting that the method above applies equipercentile equating on the discrete CDFs of the test forms, thus bypassing the continuization step to avoid favoring one method over another. This is possible because the basic requirement for finding percentiles is that the dataset can be ordered. Hence, for a set of discrete scores X, the k-th percentile will be a value P_k such that at most k % of the data are smaller in value than P_k and at most $(100 - k)$ % of the data are larger (Johnson and Kuby, 2008, Sec. 2.6). As a consequence, equivalent scores will always be integers. This is done to avoid favoring one method over another, but ends up producing equivalent scores which are always integers.

For the Beta-generated data, each test score is drawn from a random number between 0 and 1 in a $Beta(\alpha, \beta)$, which is then multiplied by the number of items in that test form. The result is a continuous, Beta-shaped probability distribution of the scores which can be cumulated to form the corresponding CDFs. This is done for both forms, and equipercentile equating is applied to the resulting CDFs to obtain $\varphi(x)$.

Unlike what happens with the real data, the procedures above allows us to obtain all the elements necessary for calculating bias and MSE as described on 2.5 and 2.6. On this study, R was set at 200, which was sufficient to give the results satisfactory consistency.

2.4 Real data application and simulation study

2.4.1 Real data application

For the real data application we used data from two administrations of the Swedish Scholastic Assessment Test (SweSAT) and the Brazilian National Assessment of Basic Education (Aneb). In both cases, the common items are administered together with the unique ones.

The SweSAT is a high-stakes, large-scale college admissions test which is given twice a year and is used for selection to higher education in Sweden. The test results are valid for five years and the examinees can retake the test as many times as they wish. Only their highest score is used when applying to colleges and universities. It is a multiple-choice, paper-and-pencil test consisting of a quantitative and a verbal section with 80 items each. The two sections are equated separately using anchor tests with 40 items each and equipercentile equating. This study only uses the quantitative section so that a unidimensional model can be used.

Aneb stands for *Avaliação Nacional da Educação Básica*, in Portuguese; it aims to gauge the overall quality of the country's school system and can thus be considered a low-stakes test from the point of view of the examinees (the students), even though it is a decisive metric for the government. Every two years since 1995, this test is administered to students from a national sample of public and private schools. Results are equated using IRTOSE between administrations and grades. This equating structure makes sense since the focus of this evaluation is on the school- and grade-level results, but microdata is publicly available at the student level, allowing us to equate the student results within a given year. Students from the 5th, 9th and 12th grades take part in Aneb. The test for 12th-graders is composed of 52 multiple-choice questions equally divided into a Math and a reading section. To obtain parallel information with the SweSAT and maintain the adequability of the chosen IRT model, this study analyzed the 26-item (13 unique, 13 common) Math test given to the 12th graders.

2.4.2 Simulation study

To the extent of our knowledge, IRT models have typically been used in the literature to randomly generate test answers even when there was no intention of fitting an IRT model afterwards. At the same time, the real world is not lacking examples of datasets which do not have the necessary assumptions for IRT modeling. We believe differences in the data-generating process can affect the performance of the equating methods. Ergo, the methods in this chapter were compared under scenarios containing data with largely different score distributions, which suggest different data-generating processes.

The simulated data is composed of four different tests: two with answers generated from randomly-drawn IRT parameters and two with answers generated from a Beta distribution. Each data-generating process contains one small test, consisting of 13 unique items and 13 common items, and one large test, with 80 unique and 40 common items. These test sizes were determined to mimic the real data study.

The simulation begins by setting $I_{\mathrm{P}} = I_{\mathrm{Q}} = I = 1\,000$ and generating a random vector of $1\,000$ examinee abilities for each we one of the two tests. Following the setup of Ogasawara (2003), we have $\theta_{\mathrm{X}+} \sim N(0, 1)$ and $\theta_{\mathrm{Y}+} \sim N(0.5, 1.2)$.

For the IRT data, we took inspiration from Andersson and Wiberg (2017) and generated a_j from an $U(0.5, 2)$ and b_j as $N(0, 1)$. Given these true item parameters as well as the previously-mentioned true skills, test answers were generated for the test forms X, Y and A.

As an attempt to create a dataset that, at least in theory, should not fit an IRT model, a second group of tests scores was generated from a Beta distribution. In particular, a $Beta(2, 5)$ was used to generate scores for test form X, a $Beta(3, 6)$ was used for Y, and a $Beta(2.5, 5.5)$ was used for A. Moreover, varying degrees of correlation between the scores on each form were determined. Specifically, the 13-item test had a correlation of 0.53 between the main test (X or Y) and the respective anchor test A; for the 80-item test, the correlation was 0.84. These values were picked to be close to those found on the real datasets of similar size: Aneb and the SweSAT had correlations around 0.54 and 0.83, respectively.

The shape parameters for the beta distributions were selected to give the test scores positive skewness, uncommon in IRT-generated data and yet present in several assessment tests in the real world, such as Aneb. Since the Beta distribution always yields values between 0 and 1, the result was multiplied by the number of items (J) in that particular test form to obtain a value between 0 and J.

Having the test scores is enough to perform KE, but to fit an IRT model, we must generate answers to each item. This is achieved by generating, for each examinee, a vector containing x_i correct answers (1) and $J_{\mathrm{X}} - x_i$ incorrect answers (0). Notice, however, that the order in which those 0s and 1s are generated will affect the latent test parameters, which will eventually be estimated by an IRT model. Simply generating, for each examinee, a sequence of 1s followed by a sequence of 0s will, once those vectors are

Table 2.1: Descriptive statistics for the real and simulated data. Bars represent averages, ρ are correlations and numbers in parenthesis correspond to standard deviations. Results for the simulated cases are averaged over 200 samples.

Statistic	SweSAT	Aneb	IRT 80	IRT 13	Beta 80	Beta 13
$\bar{X}^+$	58.4 (17.6)	8.1 (3.9)	60.7 (23.5)	12.6 (5.8)	35.4 (18.3)	7.8 (3.6)
$\bar{Y}^+$	56.5 (19.0)	8.3 (4.0)	71.9 (26.5)	13.3 (6.5)	39.2 (17.5)	8.4 (3.5)
$\bar{A}_\text{X}$	16.7 (6.4)	4.0 (2.2)	21.5 (8.1)	5.6 (3.0)	12.5 (6.2)	4.1 (2.0)
$\bar{A}_\text{Y}$	16.6 (6.6)	4.5 (2.3)	24.5 (9.1)	6.7 (3.4)	12.5 (6.2)	4.1 (2.0)
$\rho(X, A_\text{X})$	0.82	0.53	0.92	0.77	0.84	0.53
$\rho(Y, A_\text{Y})$	0.84	0.54	0.95	0.82	0.85	0.53

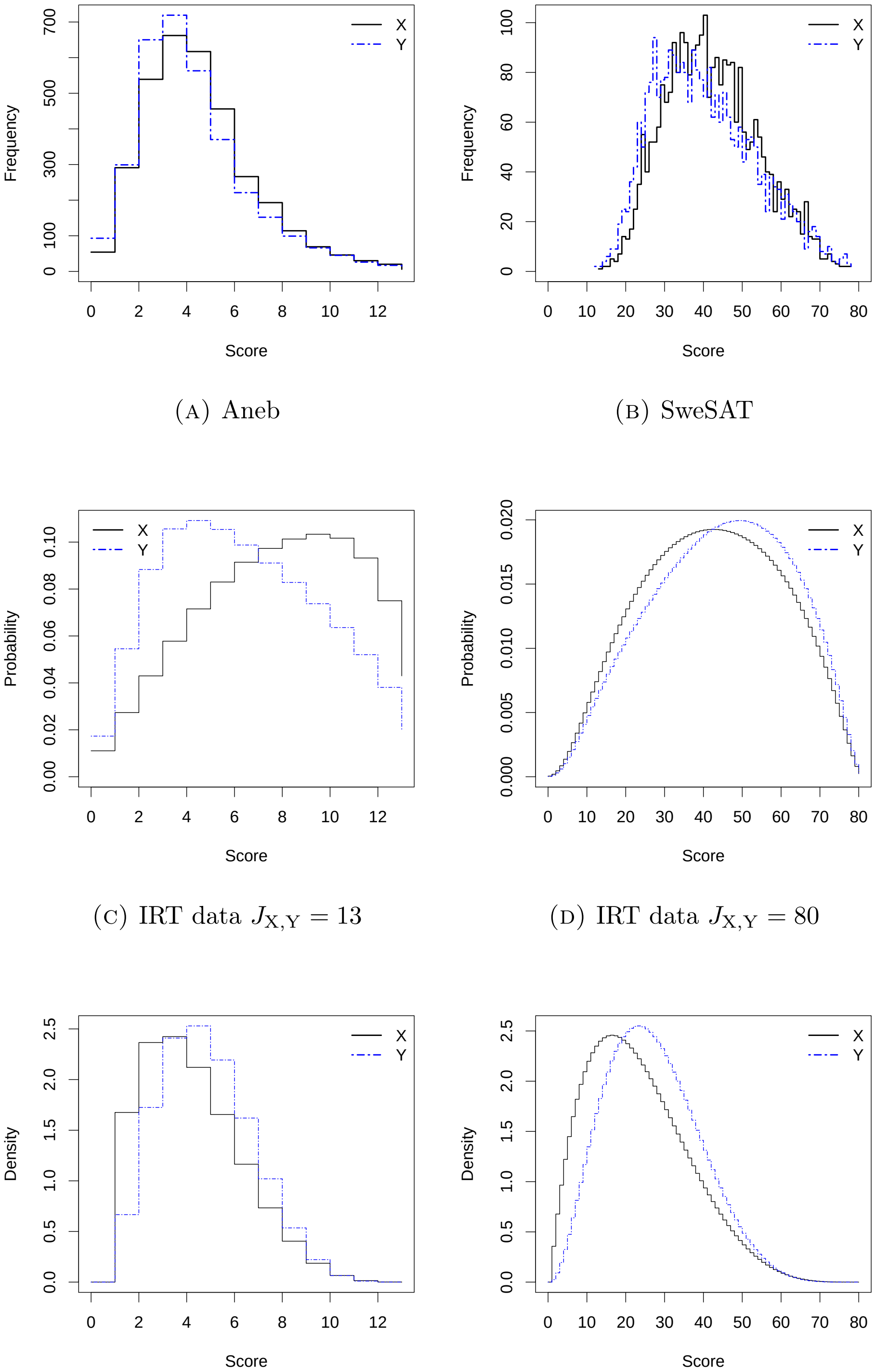
stacked to form the answer matrix for all examinees, create many items which everyone answered correctly and many others on which no one was able to get a score. On the opposite end, randomly scrambling those 0s and 1s will generate a uniform answer matrix with no items standing out as particularly easy or difficult. Both these extreme cases seem unrealistic, so a compromise was found by permutating each answer vector using hand-picked probability weights which would generate reasonable distributions of items according to difficulty.

2.5 Results

Some descriptive statistics about the real and simulated test data can be found in Tab. 2.1, where it is also shown how IRT 80 and Beta 13 were fairly successful in respectively replicating the SweSAT and the Aneb data, particularly with respect to the average scores. The other simulated cases, Beta 80 and IRT 13, present largely different average scores when compared to their similarly-sized empirical counterparts, but are nonetheless valid and realistic cases worth being analyzed.

The distribution of the observed scores for both the real and the simulated data can be seen in Fig. 2.2. The score distributions of the anchor tests were omitted, as their similarity to the respective X and Y test forms make them redundant.

The shapes of the score distributions in Fig. 2.2 can be grouped into two categories: strongly skewed distributions (Aneb and Beta), and symmetric or weakly-skewed distributions (SweSAT and IRT), with the IRT data being more platykurtic than the SweSAT. For the simulated data, examinees for test form Y have a slightly higher average ability than those for X, although that is barely visible from the plots in Fig. 2.2. This apparent similarity of distributions of X and Y can also be seen on the real data applications,



(A) Aneb

(B) SweSAT

(C) IRT data $J_{\mathrm{X,Y}} = 13$

(D) IRT data $J_{\mathrm{X,Y}} = 80$

(E) Beta data $J_{\mathrm{X,Y}} = 13$

(F) Beta data $J_{\mathrm{X,Y}} = 80$

FIGURE 2.2: Distribution of observed test scores. For simulated data, values correspond to the theoretical score probabilities/densities.

Table 2.2: Bias/error per score for the simulation study and real data application with 13 items.

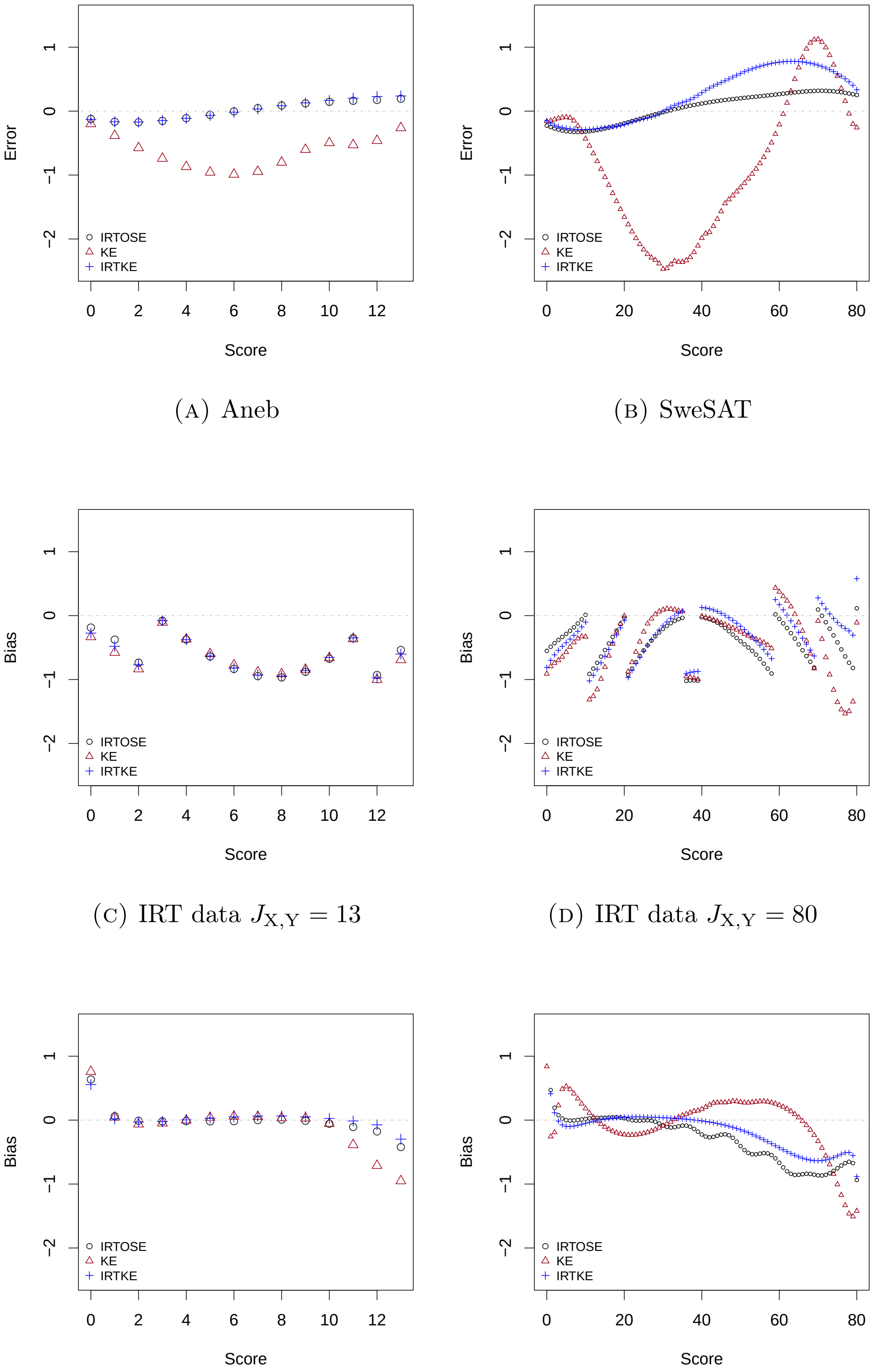
Score	Aneb			IRT $J_{\mathrm{X,Y}} = 13$			Beta $J_{\mathrm{X,Y}} = 13$		
	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE
0	-0.122	-0.199	-0.128	-0.187	-0.333	-0.274	0.635	0.757	0.555
1	-0.165	-0.384	-0.166	-0.376	-0.578	-0.479	0.060	0.041	0.020
3	-0.152	-0.741	-0.148	-0.079	-0.107	-0.079	-0.018	-0.048	-0.021
5	-0.057	-0.957	-0.066	-0.639	-0.598	-0.631	-0.020	0.039	0.030
7	0.047	-0.945	0.034	-0.947	-0.885	-0.930	-0.001	0.054	0.064
9	0.122	-0.602	0.129	-0.878	-0.842	-0.856	-0.010	0.030	0.052
11	0.162	-0.529	0.202	-0.351	-0.365	-0.349	-0.108	-0.387	-0.013
12	0.176	-0.462	0.228	-0.928	-1.002	-0.974	-0.177	-0.712	-0.074
13	0.194	-0.262	0.238	-0.537	-0.692	-0.602	-0.421	-0.952	-0.298
Avg. abs. bias	0.123	0.629	0.135	0.608	0.639	0.624	0.111	0.231	0.093
Avg. MSE	0.018	0.459	0.022	0.463	0.498	0.476	0.085	0.313	0.074

perhaps with the exception of the SweSAT, where test form Y seems to have a slightly lower average than test form X.

Information regarding the quality of the equating transformations can be seen in Figs. 2.3 and 2.4, respectively presenting the bias/error and MSE per score, data and method. These statistics were calculated according to Eqs. (2.5) and (2.6). The figures show similar patterns for IRTOSE and IRTKE, which is expected since these methods only differ by their continuization algorithm as well as by how IRTOSE transforms the item parameter estimates of the non-base forms—as shown in Fig. 1.3 and further detailed in Sec. 3.1.1 ahead—, whereas IRTKE does not use them under CE. On the other hand, the behavior of KE presents itself with more distinction, particularly on Aneb, the SweSAT and the Beta-generated data with 80 items.

Tables 2.2 and 2.3 supplement the graphical information provided by Fig. 2.3. They summarize the numerical results for the bias/errors, facilitating comparisons between data with the same number of items. Some scores on those tables were omitted for brevity, but were still included in the averages. Moreover, they include the average absolute bias and the average MSE, which aid in the comparison of the equating methods.

Corroborating what was observed in Fig. 2.3 and using the absolute value of the average bias as the evaluation criterion, the tables show IRTOSE and IRTKE performing slightly better than KE in all but one of the scenarios studied. For Beta 80, KE offered less average bias than IRTOSE, but it did so at the cost of more average MSE. Nonetheless, it must be noted that the lack of a significance threshold on this kind of evaluation criteria paired with how the results from these tables often differ by less than one unit make it difficult to point out a clear, universal winner.



(A) Aneb

(B) SweSAT

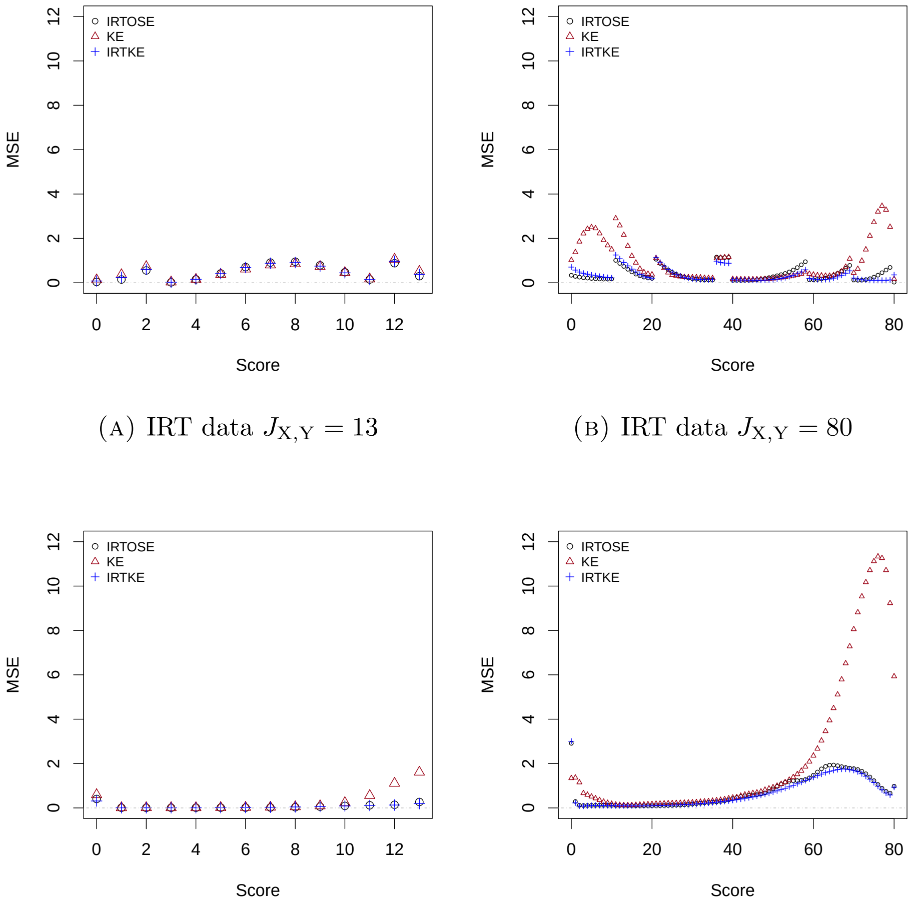
(C) IRT data $J_{X,Y} = 13$

(D) IRT data $J_{X,Y} = 80$

(E) Beta data $J_{X,Y} = 13$

(F) Beta data $J_{X,Y} = 80$

FIGURE 2.3: Bias/error per score.



(A) IRT data $J_{\mathrm{X,Y}} = 13$

(B) IRT data $J_{\mathrm{X,Y}} = 80$

(C) Beta data $J_{\mathrm{X,Y}} = 13$

(D) Beta data $J_{\mathrm{X,Y}} = 80$

FIGURE 2.4: MSE per score.

TABLE 2.3: Bias/error per score for the simulation study and real data application with 80 items.

Score	SweSAT			IRT $J_{\mathrm{X,Y}} = 80$			Beta $J_{\mathrm{X,Y}} = 80$		
	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE
0	-0.226	-0.159	-0.151	-0.551	-0.910	-0.810	1.698	0.839	1.723
10	-0.321	-0.431	-0.287	0.012	-0.331	-0.102	0.019	0.182	-0.049
20	-0.191	-1.656	-0.207	-0.034	-0.005	-0.078	0.029	-0.226	0.046
30	-0.017	-2.469	-0.006	-0.208	0.091	-0.158	-0.084	-0.090	0.038
40	0.119	-1.986	0.288	-0.026	-0.009	0.126	-0.225	0.171	-0.013
50	0.202	-1.194	0.590	-0.400	-0.252	-0.167	-0.407	0.281	-0.151
60	0.266	-0.210	0.766	-0.051	0.372	0.172	-0.668	0.234	-0.432
70	0.318	1.122	0.719	0.095	-0.084	0.278	-0.864	-0.329	-0.636
80	0.252	-0.257	0.335	0.114	-0.110	0.579	-0.936	-1.422	-0.883
Avg. abs. bias	0.215	1.138	0.402	0.408	0.486	0.361	0.353	0.316	0.237
Avg. MSE	0.055	1.903	0.221	0.381	0.929	0.351	0.715	2.291	0.648

On an Intel i5-2467M CPU with 4 GB of RAM, the IRT-simulated data took around 5 seconds to perform IRTOSE, 1 s to perform KE and 6 s to perform IRTKE on the 13-item tests. For the larger test, those times were around 65 s, 22 s and 83 s, respectively. Time accounted for all the steps depicted in Fig. 2.1. Numbers were similar for the Beta-generated data, which suggest little to no extra overhead to fitting an IRT model to highly-skewed data. For the SweSAT, respective runtimes for IRTOSE, KE and IRTKE were 85 s, 28 s and 103 s; for Aneb, they clocked around 10 s, 2 s and 12 s. These results are within the expected values, given the size of each dataset as well as the particularities of each equating method.

2.6 Discussion

This chapter expands on the work of Wiberg and González (2016), achieving its aim of suggesting a method of comparing equating transformations from different frameworks. It also advances the work of Leôncio and Wiberg (2018) by introducing and implementing a framework for conducting computer simulations to evaluate the equating transformations under study. In addition to traditional IRT methods to generate data, this study derived a method to generate item responses without relying on IRT parameters so that comparisons between IRT and non-IRT equating methods can be done on fair grounds.

Most of the observed scenarios suggested that IRTOSE and IRTKE outperform KE with respect to bias and MSE. Half of those scenarios have results that favor IRTOSE over IRTKE, but the differences are often so small that the results could have been the opposite under slightly different conditions.

These results are in line with those from Meng (2012), for whom IRT equating showed more accuracy than KE, even though he also observed that KE shows more stability than IRTOSE, which we could not confirm in our scenarios. Our preliminary tests have indicated, however, that much of the performance of KE seems to rely on how well the analyst can find a reasonable pre-smoothing model for the observed score distributions, particularly when the score distribution contains scores with few observations. KE can also be proven quite useful when speed is a priority, since it does not require calculations at the item level and can offer results at a fraction of time compared with IRT methods. Moreover, even though IRT-based equating methods may not suffer from as high a degree of dependence on model fit as KE, they do require more assumptions to be implemented. As a matter of fact, it is pointless to try to fit a model if the data does not meet a framework's basic assumptions in the first place. Thus, KE may offer a suitable alternative on instances where IRT frameworks should not be applied.

The results can also be studied by grouping the scenarios according to two characteristics: skewness of the score distribution and test size. Before confronting results between real data and simulated data, it is important to keep in mind that, unlike the single samples provided by each of the real datasets, the simulated data results are composed of an average over hundreds of replications, which gives more stability and confidence to the results.

IRTOSE and IRTKE performed well both in fairly symmetric and asymmetric score distributions. KE, on the other hand, seems to have its performance mostly bound to the quality of the pre-smoothing model.

The continuity breaks observed on the bias and MSE curves for the IRT data—which are present in both test sizes but are more evident in the 80-items case—are caused by our decision to have all $\varphi(x)$ be integers for the IRT-generated data. Specifically, the breaks coincide with the scores where the magnitude of the difference between the equated scores changes. Any method of continuization could smooth those breaks, but would have favored a particular framework.

The comparison between short tests and long tests does not seem to suggest a correlation between test size and error, with all scenarios containing most of the bias between -1 and 1.

It is important to keep in mind that the outcome of all the equating methods studied is a result of several choices of models and parameters. Depending on the decisions taken at each step of the equating, we can observe variations in the output that could ultimately turn the decision in favor a particular method over another. The discussion about how to create the best environment possible to allow fair comparisons remains open, but we believe our contribution has helped shed some light into the debate.

An important characteristic of a fair comparison is its independence from subjectivity. This motivated our decision to take a hands-free approach to pre-smoothing, in which a stepwise procedure chose a model instead of having a human manually checking the goodness-of-fit of countless model for each of our hundreds of samples. The day-to-day usage, however, often contains only one dataset and several methods to choose from. Under these conditions, we recommend careful experimentation and observation of the sensitivity of the results to the different alternatives. This attention is especially important when dealing with high-stakes tests such as admissions tests, where the choice of a particular method can mean the difference between accepting an examinee into a university program or not.

Regarding the construction of a test booklet, we believe that focus should be put on the quality of the test items instead of their quantity. This is especially important if IRT models are expected to be used, although the efficacy of pre-smoothing methods for KE can also be harmed by the presence of items that are too difficult, easy, tricky or confusing.

Further studies should focus not only on the application of the methods developed here to other data and with other simulation and equating parameters but also its generalization to more than two test forms, internal anchor items and equating frameworks such as those mentioned in Section 2.1. In particular, it would be interesting to see the development of other methods for working with real data; the method applied here was pointed out by Harris and Crouse (1993) as having some shortcomings such as the dependency on which form was taken as the base, but they still take it as useful for checking the adequacy of an equating method or data collection design.

Some authors tend to refer to the equating procedure used here—equating a test form to itself—as "circular equating", but we chose to avoid this nomenclature due to it not being universal. For example, Wang *et al.* (2000) found circular equating to be generally invalid for evaluating the adequacy of equating. However, they define circular equating as "equating a test for to itself through a chain of equating". The case studied in this chapter differs from the ones studied in Wang *et al.* (2000) due to the absence of a chained structure, so their conclusions might not apply to the case under study but do instigate further investigation into better ways to generate expected equated score distributions for empirical data.

We chose well-known summary indices to evaluate the equating transformations, which leaves the creation of suitable equating-specific measures for future studies. For instance, the DTM can be easily applied to bias calculations, but more complex indices like the PRE might need adjustments before they can be applied to cross-framework comparisons. Future studies of these summary indices could also address other issues pointed out by Harris and Crouse (1993), such as the choice of the associated loss function.

Finally, we highlight the importance of further research comparing different equating frameworks, even if it is unlikely that an unambiguous choice will surface from such studies (Kolen and Brennan, 2014). Future studies should also analyse the sensitivity of the results to changes in the number of items and replications of each simulation scenario.

Chapter 3

A likelihood approach to IRT equating of multiple forms

3.1 Introduction

Let us consider an achievement test being administered using the NEAT (non-equivalent groups with anchor test, von Davier *et al.*, 2004) data-collection design. To estimate the ability of the subjects under examination, IRT models are fit to the test data generated by those examinees. Often, different models are independently fit to each test form. This usually happens when each form is administered in a different point in time, but can also occur in simultaneously-delivered forms, due to the advantages of separate item calibration when compared with simultaneous calibration detailed below.

When fitting an IRT model to test data, the most common estimation method used is the maximization of the marginal likelihood function (Bock and Aitkin, 1981), which assumes the abilities to have a Normal distribution. This is a problem whenever the groups taking each form are not equivalent, which is an assumption of the NEAT data collection design. There are two common ways to solve this issue:

1. observe the differences in the parameter estimates for the items in common between two forms and replicate those differences to the rest of the items;

2. (re-)estimate all the parameters for all the forms simultaneously.

At first sight, the first alternative—called "separate calibration"—may sound cumbersome, with the second—known as "concurrent calibration"—being the most sensible solution. However, as Kolen and Brennan (2014) have pointed out, separate calibration seems to be the safest option of the two. Concurrent calibration might indeed provide better IRT estimates under certain situations, but separate calibration is more robust to IRT violations. Moreover, concurrent calibration tends to provide less accurate results when there are few common items (Kim and Cohen, 1998) and might be computationally unfeasible if there are so many test forms that the aggregate number of items involved in equating is in the thousands (Haberman, 2009).

3.1.1 Parameter transformation in the 2PL

From a statistical point of view, the problem of independent IRT scales introduced by separate calibration is solved by realizing that the probability of correctly answering an item in the 2PL model is invariant to linear transformations of the parameters (Bartolucci *et al.*, 2016, Sec. 3.4). In other words, if one particular 2PL model fits the data, then any linear transformation of the parameters in that model will fit that data just as well.

Let the following two-parameter logistic model (2PL) be used to estimate the probability that individual i taking test t correctly answers a dichotomous item j given the person's ability ($\theta_{it} \in \mathbb{R}$) and the item's discrimination ($a_{jt} > 0$) and difficulty ($b_{jt} \in \mathbb{R}$). Algebraically, we have

$$\Pr(X_{ijt} = 1|\theta_{it}; a_{jt}, b_{jt}) = \frac{1}{1 + \exp\left[-a_{jt}(\theta_{it} - b_{jt})\right]}. \tag{3.1}$$

Before we can proceed with transforming item parameters across test forms, we must set some constraints to θ. This is necessary because the IRT model above will otherwise not be identifiable. To prove this, let us consider the following linear transformations:

$$\begin{aligned} \theta_{i\mathrm{Y}} &= A\theta_{i\mathrm{X}} + B \\ b_{j\mathrm{Y}} &= Ab_{j\mathrm{X}} + B \\ a_{j\mathrm{Y}} &= a_{j\mathrm{X}}/A \end{aligned} \tag{3.2}$$

Considering that jX and jY correspond to the same item j that is common to both test forms X and Y, we use the parameters above on Eq. (3.1) and obtain

$$\begin{aligned}
\Pr(X_{ij\mathrm{Y}} = 1|\theta_{i\mathrm{Y}}; a_{j\mathrm{Y}}, b_{j\mathrm{Y}}) &= \frac{1}{1+\exp\left[-a_{j\mathrm{Y}}(\theta_{i\mathrm{Y}} - b_{j\mathrm{Y}})\right]} \\
&= \frac{1}{1+\exp\left[-\frac{a_{j\mathrm{X}}}{A}(A\theta_{i\mathrm{X}} + B - Ab_{j\mathrm{X}} - B)\right]} \\
&= \frac{1}{1+\exp\left[-a_{j\mathrm{X}}(\theta_{i\mathrm{X}} - b_{j\mathrm{X}})\right]} \\
&= \Pr(X_{ij\mathrm{X}} = 1|\theta_{i\mathrm{X}}; a_{j\mathrm{X}}, b_{j\mathrm{X}}).
\end{aligned}$$

This means that both $\{\theta_{i\mathrm{X}}, a_{j\mathrm{X}}, b_{j\mathrm{X}}\}$ and $\{\theta_{i\mathrm{Y}}, a_{j\mathrm{Y}}, b_{j\mathrm{Y}}\}$ give the same probability of correctly answering item j, and the result is the same for any value of A and B. In other words, there are infinite parameters that will give two people with different ability levels the same probability to correctly answer an item, thus making this model unidentifiable.

Let us now impose a couple of constraints to the latent trait θ. Specifically, let $\theta \sim N(0,1)$, which means that $\mu_\theta = 0$ and $\sigma_\theta = 1$. Not only is this an obvious way to remove the indeterminacy of the θ scale, but also the only reason for the necessity of eventually transforming the person and item parameters (van der Linden and Barrett, 2016). After all, if X and Y are test forms administered to samples from potentially different populations, they will have their own distribution parameters of the latent trait instead of the common $N(0,1)$ metric. Nonetheless, conversion to and from the standard Normal distribution is trivial. The implications of such standardization on the item parameter estimates, however, needs to be further examined.

Consider the part of the denominator of Eq. (3.1) which contains the item and person parameters, i.e., $a_{jt}(\theta_t - b_{jt})$ (the person index i was dropped here for visual simplification, but does not influence the result of the operations). For $t = \mathrm{X}$, we perform the following operation:

$$\begin{aligned}
a_{j\mathrm{X}}(\theta_\mathrm{X} - b_{j\mathrm{X}}) &= \frac{a_{j\mathrm{X}}\sigma_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}}\left(\theta_\mathrm{X} - b_{j\mathrm{X}} + \mu_{\theta_\mathrm{X}} - \mu_{\theta_\mathrm{X}}\right) \\
&= a_{j\mathrm{X}}\sigma_{\theta_\mathrm{X}}\left(\frac{\theta_\mathrm{X} - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}} - \frac{b_{j\mathrm{X}} - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}}\right).
\end{aligned} \tag{3.3}$$

This shows that when the abilities are standardized so that θ_X has mean zero and unitary standard deviation, the item parameters are also transformed, with a_{jt} and b_{jt} respectively becoming $a_{j\mathrm{X}}\sigma_{\theta_\mathrm{X}}$ and $(b_{j\mathrm{X}} - \mu_{\theta_\mathrm{X}})/\sigma_{\theta_\mathrm{X}}$. The same will happen to form Y when standardizing θ_Y, yielding $a_{j\mathrm{Y}}\sigma_{\theta_\mathrm{Y}}$ and $(b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{Y}})/\sigma_{\theta_\mathrm{Y}}$. Together, these transformations allow us to convert the item parameters of one form directly to the scale of the other.

For example, if we were to go from form Y to form X, we would like the item difficulties in Y to be transformed like those at the end of Eq. (3.3), i.e.,

$$\frac{b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}}.$$

In other words, the item difficulties in form Y are being standardized using the mean and standard deviation of the abilities in form X. Going backwards from the desired result, consider the following operation:

$$\begin{aligned}\frac{b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}} &= \frac{\left(\frac{b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{Y}}}{\sigma_{\theta_\mathrm{Y}}}\sigma_{\theta_\mathrm{Y}} + \mu_{\theta_\mathrm{Y}}\right) - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}} \\ &= \frac{b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{Y}}}{\sigma_{\theta_\mathrm{Y}}}\frac{\sigma_{\theta_\mathrm{Y}}}{\sigma_{\theta_\mathrm{X}}} + \frac{\mu_{\theta_\mathrm{Y}} - \mu_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{X}}} \\ &= \frac{b_{j\mathrm{Y}} - \mu_{\theta_\mathrm{Y}}}{\sigma_{\theta_\mathrm{Y}}}A_\mathrm{YX} + B_\mathrm{YX}.\end{aligned} \tag{3.4}$$

We let $A_\mathrm{YX} = \sigma_{\theta_\mathrm{Y}}/\sigma_{\theta_\mathrm{X}}$ and $B_\mathrm{YX} = (\mu_{\theta_\mathrm{Y}} - \mu_{\theta_\mathrm{X}})/\sigma_{\theta_\mathrm{X}}$ above so that we could define A_YX and B_YX as the equating coefficients that convert the item parameters from the scale of form Y to that of form X.

Conveniently, by transforming the difficulty parameters we also obtain everything we need to convert the item discriminations: in this example, we want the parameters $a_{j\mathrm{Y}}$ to be on the scale of form X, which means performing the following transformation:

$$a_{j\mathrm{Y}}\sigma_{\theta_\mathrm{X}} = a_{j\mathrm{Y}}\sigma_{\theta_\mathrm{X}}\frac{\sigma_{\theta_\mathrm{Y}}}{\sigma_{\theta_\mathrm{Y}}} = a_{j\mathrm{Y}}\sigma_{\theta_\mathrm{Y}}\frac{\sigma_{\theta_\mathrm{X}}}{\sigma_{\theta_\mathrm{Y}}} = a_{j\mathrm{Y}}\sigma_{\theta_\mathrm{Y}}\frac{1}{A_\mathrm{YX}}.$$

The equating coefficients derived above correspond to the A and B parameters in the transformations from (3.2). Perhaps more important than transforming the item parameters, we can now transform the abilities from one form to another, thus allowing the direct comparison of abilities of two groups who took different versions of a test. The equating procedure is complete at this point, but for practical purposes one can proceed to convert abilities into test scores, which are usually bounded and non-negative, therefore easier for a layman to interpret. Since this extra step is performed after the estimation of the equating coefficients, it will not be considered in this chapter.

3.1.2 Pairwise and multiple form transformation methods

In practical applications, the equating coefficients cannot be directly calculated because their building blocks—the means and variances of the abilities of each group—are unknown. However, they can be estimated using the data available from the test administrations. Transformation methods for equating a pair of test forms include moment-based approaches like the mean-sigma (Marco, 1977) and the mean-mean (Loyd and Hoover, 1980) as well as solutions based on the item characteristic curves such as the Haebara (Haebara, 1980) and the Stocking–Lord (Stocking and Lord, 1982) methods.

These methods are very useful for equating a pair of test forms, but they do not correctly generalize to the scenario where more than two forms are to be equated. Theoretically, one could chain pairwise equating transformations to equate a large number of forms, but this would mean that only one of the links of a particular form will be taken into account when calculating the equating parameters. To equate multiple forms, it is better to deploy methods that take into account the linkage plan as a whole, calculating all equating coefficients simultaneously and considering all the links between the test forms.

Multiple equating methods are not new in the literature, with Haberman (2009) proposing a linear regression method, Battauz (2013) presenting chain and average equating coefficients and Battauz (2017a) introducing the generalization of some well-known pairwise methods such as those mentioned above. One basic difference between pairwise and multiple form transformation methods concerns the calculation of the equating coefficients. Instead of defining A and B as per Eq. (3.4), which always involves parameters from two forms, one must now consider all forms simultaneously.

In practice, the estimation of the equating coefficients A_t and B_t for a certain form t will involve the item parameters for that form—(a_{jt}, b_{jt})—as well as a set of item parameters (a_j^*, b_j^*) that represent a common metric for item j across all test forms. These relations are illustrated on Eq. (3.5) below:

$$\begin{aligned} a_j^* &= a_{jt}/A_t \\ b_j^* &= A_t b_{jt} + B_t \end{aligned} . \tag{3.5}$$

As meaningful as the contributions from the works by Haberman (2009); Battauz (2017a) were, they rely on a couple of assumptions that should be addressed to amplify the applicability of multiple-form equating methods: they assume independence between the item parameter estimates and homoskedasticity of those estimates, when parameter estimates of the same item are actually always correlated and heteroskedasticity can be expected in tests containing some items with extreme parameter values. The next section presents a new method for multiple-form equating that addresses these issues.

3.2 Methodology of likelihood equating

According to Reise and Revicki (2015), the non-linearity of IRT models imply that direct, analytical solutions are impossible, and iterative algorithms must be employed. Bartolucci *et al.* (2016) list three main approaches to estimate IRT parameters: the conditional maximum likelihood (CML), the joint maximum likelihood (JML), and the marginal maximum likelihood (MML).

The CML is restricted to Rasch-type models, so it is of no use to the 2PL model family considered in this thesis. As for JML and MML, the most important difference between them is how each one treats the latent abilities θ. In JML, θ is considered a parameter, just like the item parameters a, b and c. In effect, the JML—and the CML, for that matter—formulates the IRT model as fixed effects. MML, conversely, treats θ as a random variable, thus making the 2PL a random-effects model (Bartolucci *et al.*, 2016). In practice, the MML procedure begins with the specification of a probability distribution for θ—usually standard Normal, as mentioned before—followed by the estimation of the item parameters given the distribution of θ.

On this thesis, we assume estimation is always done using MML, due to its estimators being consistent, unlike those provided by JML (Bock and Aitkin, 1981; van der Linden and Hambleton, 1998).

3.2.1 The likelihood function

Let (a_{jt}, b_{jt}) be the true item parameters of item j in form t and $(\hat{a}_{jt}, \hat{b}_{jt})$ be their estimates obtained through MML. From Eq. (3.5), we know that $a_{jt} = A_t a_j^*$ and $b_{jt} = (b_j^* - B_t)/A_t$. Hence, their estimates can be modeled as

$$\begin{aligned} \hat{a}_{jt} &= a_{jt} + \epsilon_{jt}^a = A_t a_j^* + \epsilon_{jt}^a \\ \hat{b}_{jt} &= b_{jt} + \epsilon_{jt}^b = \frac{b_j^* - B_t}{A_t} + \epsilon_{jt}^b. \end{aligned} \tag{3.6}$$

Since the item parameters are estimated using the maximum likelihood method, we know that both ϵ_{jt}^{a} and ϵ_{jt}^{b} are asymptotically-distributed as a zero-mean Normal, and the item parameter estimates are also normally-distributed as follows:

$$\begin{aligned} \hat{a}_{jt} &\overset{a}{\sim} N\left(A_t a_j^*, \sigma_{a_{jt}}^2\right) \\ \hat{b}_{jt} &\overset{a}{\sim} N\left(\frac{b_j^* - B_t}{A_t}, \sigma_{b_{jt}}^2\right). \end{aligned} \tag{3.7}$$

We begin by assuming the simplest case of independence and homoskedasticity of the item parameter estimates. Explicitly, the independence condition gives us $f(\hat{a}_{jt}, \hat{b}_{jt}) = f(\hat{a}_{jt})f(\hat{b}_{jt})$, where $f(\cdot)$ is the probability density function (PDF) of the Normal distribution and $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are vectors of all item parameters across all forms. For homoskedasticity, we have that $\sigma_{a_{jt}}^2 = \sigma_a^2$ and $\sigma_{b_{jt}}^2 = \sigma_b^2$ for all $j = 1, \ldots, J_t$ and $t = 1, \ldots, T$. For convenience, the range of test forms is hereupon represented by numbers instead of letters. Altogether, the likelihood function under these conditions is given by

$$L\left(\mathbf{A}, \mathbf{B}, \mathbf{a}^*, \mathbf{b}^*, \sigma_a^2, \sigma_b^2; \hat{\mathbf{a}}, \hat{\mathbf{b}}\right) = \prod_{t=1}^{T}\prod_{j=1}^{J_t} f(\hat{a}_{jt})f(\hat{b}_{jt}), \tag{3.8}$$

where T is the total number of test forms to be equated and J_t is the number of items in form t. $\mathbf{A}$ and $\mathbf{B}$ are vectors respectively containing A_t and B_t for all t; $\mathbf{a}^*$, $\mathbf{b}^*$, $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are also vectors containing the similarly-named objects referenced in Eq. (3.7). The maximization of L yields the estimates of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{a}^*$, $\mathbf{b}^*$, σ_a^2, and σ_b^2.

3.2.2 The profile likelihood function

When dealing with large-scale assessments, often composed of several test forms with dozens or hundreds of items each, the number of parameters to be estimated can quickly become a concern. After all, each new item adds at least one IRT parameter to the likelihood function (two in the 2PL model), and any additional test form can introduce several new items as well as two mandatory equating coefficients. This can easily make using the full likelihood too overwhelming from a computational perspective.

One way to overcome this problem is by using the profile likelihood function to reduce the number of parameters to be estimated. Fortunately, the case under study is a great candidate for this type of procedure. In effect, out of all the parameters to be estimated in the likelihood function depicted in Eq. (3.8), the equating coefficients A_t and B_t are the ones of utmost interest. Arguably, the item parameter variances $\sigma_{a_{jt}}^2$ and $\sigma_{b_{jt}}^2$ can also be considered of interest, due to their usefulness in practical applications as indicators of potential problems in the IRT model fitting which could propagate into the equating procedures. All other parameters—namely, $\mathbf{a}^*$ and $\mathbf{b}^*$—can be considered nuisance parameters. It is of great interest to eliminate these from the optimization procedure, given their potentially large size.

In order to calculate the profile likelihood of the equating coefficients $\mathbf{A}$ and $\mathbf{B}$, first we need to find the maximum likelihood estimators (MLE) of the nuisance parameters. The step-by-step calculations are presented on Appendix A, and the final results for the independent and homoskedasticity case are

$$\hat{a}^*_j = \frac{\sum_t \frac{\hat{a}_{jt}}{A_t} A_t^2}{\sum_t A_t^2} \qquad \hat{b}^*_j = \frac{\sum_t \frac{\hat{b}_{jt} A_t + B_t}{A_t^2}}{\sum_t \frac{1}{A_t^2}}. \tag{3.9}$$

Once the nuisance parameters are defined as functions of the parameters of interest and the data, the vectors of a^*_j and b^*_j can be replaced by their MLEs, thus reducing the number of parameters to be estimated by $2J$. As a result, instead of working with $L(\mathbf{A}, \mathbf{B}, \mathbf{a}^*, \mathbf{b}^*, \sigma_a^2, \sigma_b^2; \hat{\mathbf{a}}, \hat{\mathbf{b}})$, we can now use the more parsimonious $L_p(\mathbf{A}, \mathbf{B}, \sigma_a^2, \sigma_b^2; \hat{\mathbf{a}}, \hat{\mathbf{b}})$.

3.2.3 Accounting for dependence and heteroskedasticity

The likelihood function introduced in Eq. (3.8) assumes the independence between $\hat{a}_{jt}$ and $\hat{b}_{jt}$ for all $j = 1, \ldots, J_t$ and $t = 1, \ldots, T$. The same goes for the MLEs presented in Eq. (3.9) above. Moreover, it was then assumed that the variances of the item parameter estimates were homoskedastic, i.e., $\sigma^2_{a_{jt}} = \sigma^2_a$ and $\sigma^2_{b_{jt}} = \sigma^2_b$. Let us now consider the consequences of assuming the existence of correlations between the item parameter estimates as well as the possibility of item-individualized variances.

If there is correlation between the item parameters, one must consider the joint distribution of $\hat{a}_{jt}$ and $\hat{b}_{jt}$, as opposed to what was defined on Eq. (3.7). Specifically, now we have to assume a joint distribution for the $2J_t$ vector of all item parameter estimates in form t, i.e.,

$$\begin{pmatrix} \hat{\mathbf{a}}_t \\ \hat{\mathbf{b}}_t \end{pmatrix} \overset{a}{\sim} N\left(\begin{pmatrix} \mathbf{a}_t \\ \mathbf{b}_t \end{pmatrix}, \boldsymbol{\Sigma}_t \right). \tag{3.10}$$

This joint distribution is contained within one test form t because the parameters for different items pertaining to separate forms are assumed to be independent. Here, $\hat{\mathbf{a}}_t$ is the vector of all item discrimination estimates $\hat{a}_{jt}$ for one form t, $\hat{\mathbf{b}}_t$ is the analogous vector for the difficulty parameters and $\boldsymbol{\Sigma}_t$ is the covariance matrix. The likelihood function in this case differs from the one in Eq. (3.8) and is now formulated as

$$L\left(\mathbf{A}, \mathbf{B}, \mathbf{a}^*, \mathbf{b}^*, \sigma^2_{a_{jt}}, \sigma^2_{b_{jt}}; \hat{\mathbf{a}}, \hat{\mathbf{b}}\right) = \prod_{t=1}^{T} f(\hat{\mathbf{a}}_t, \hat{\mathbf{b}}_t), \tag{3.11}$$

where $f(\cdot)$ is the PDF of the multivariate Normal distribution with parameters given in Eq. (3.10). Akin to the case depicted in Eq. (3.7), the parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{a}^*$, $\mathbf{b}^*$, $\sigma^2_{a_{jt}}$ and $\sigma^2_{b_{jt}}$ enter the equation as part of the distribution parameters (mean vector and covariate matrix) of the underlying probability distribution of $\hat{\mathbf{a}}_t$ and $\hat{\mathbf{b}}_t$.

To estimate the equating coefficients that maximize the likelihood function in this new scenario, once again we are faced with the issue of dealing with nuisance parameters. However, we can no longer use the MLEs from Eq. (3.9) to generate a profile likelihood where $\mathbf{a}^*$ and $\mathbf{b}^*$ are no longer part of the equation; those MLEs were calculated assuming independence between the item parameters, which is no longer an assumption hereupon. Moreover, the presence of a potentially-dense $\mathbf{\Sigma}_t$ means that a Generalized Least Squares (GLS, Aitkin, 1935) approach would be more appropriate to define an estimator for the nuisance parameters $\mathbf{a}^*$ and $\mathbf{b}^*$ as part of the maximum profile likelihood procedure for estimating the equating coefficients in Eq. (3.11). This requires the reformulation of the equations in (3.6), which gives us

$$\begin{cases} \frac{1}{A_t}\hat{\mathbf{a}}_t = \mathbf{a}^* + \frac{1}{A_t}\epsilon^{\mathbf{a}}_{\mathbf{t}} \\ A_t\hat{\mathbf{b}}_t + B_t = \mathbf{b}^* + A_t\epsilon^{\mathbf{b}}_{\mathbf{t}} \end{cases}.$$

The equations in the system above jointly have the form $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{Y}$, β and ϵ vectors are juxtaposing the elements from both rows in the equation above. In extended form, the model will look as follows:

$$
\begin{array}{ccccccc}
\mathbf{Y} & = & \mathbf{X} & \cdot & \beta & + & \epsilon \\
\begin{pmatrix}
\hat{a}_{11}/A_1 \\ \vdots \\ \hat{a}_{J_1 1}/A_1 \\ \vdots \\ \hat{a}_{1T}/A_T \\ \vdots \\ \hat{a}_{J_T T}/A_T \\ A_1\hat{b}_{11}+B_1 \\ \vdots \\ A_1\hat{b}_{J_1 1}+B_1 \\ \vdots \\ A_T\hat{b}_{1T}+B_T \\ \vdots \\ A_T\hat{b}_{J_T T}+B_T
\end{pmatrix}
& = &
\begin{pmatrix}
1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 1
\end{pmatrix}
& \cdot &
\begin{pmatrix}
a_1^* \\ \vdots \\ a_J^* \\ b_1^* \\ \vdots \\ b_J^*
\end{pmatrix}
& + &
\begin{pmatrix}
\epsilon^a_{11}/A_1 \\ \vdots \\ \epsilon^a_{J_1 1}/A_1 \\ \vdots \\ \epsilon^a_{JT}/A_T \\ \vdots \\ \epsilon^a_{J_T T}/A_T \\ A_1\epsilon^b_{11} \\ \vdots \\ A_1\epsilon^b_{J_1 1} \\ \vdots \\ A_T\epsilon^b_{J_T T} \\ \vdots \\ A_T\epsilon^b_{J_T T}
\end{pmatrix}
\end{array}
$$

With respect to their sizes, $\mathbf{Y}$ and ϵ are vectors with $2\sum_t J_t$ elements. β has length $2J$ and $\mathbf{X}$ is a $2\sum_t J_t \times 2J$ binary design matrix. In accordance with what was explained in Sec. 1.2, page 4, due to the existence of common items between forms, the number of different items administered in the test, J, is smaller than the sum of the number of items administered across all forms, $\sum_t J_t$.

The GLS procedure allows the estimation of β as a function of the design matrix, the response vector $\mathbf{Y}$ and the covariance matrix of the regression residuals $\mathbf{\Omega}$ as

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{Y}. \tag{3.12}$$

Since only items being administered on the same form are correlated, i.e., $\rho(a_{jt}b_{j't'}) = 0$ for all $t \neq t'$, the covariances between the residuals of those items parameter estimates are also zero. We can then represent $\mathbf{\Omega}$ as the following block diagonal matrix:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_T \end{bmatrix}_{2\sum_t J_t \times 2\sum_t J_t} .$$

Here, the blocks contain the covariance matrices of the residuals of the parameter estimates for items belonging to the same test form. Explicitly:

$$\mathbf{\Omega}_t = \begin{bmatrix} \sigma_{\epsilon_{a_1}\epsilon_{a_1}}/A_t^2 & \cdots & \sigma_{\epsilon_{a_1}\epsilon_{a_{J_t}}}/A_t^2 & \sigma_{\epsilon_{a_1}\epsilon_{b_1}} & \cdots & \sigma_{\epsilon_{a_1}\epsilon_{b_{J_t}}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\epsilon_{a_{J_t}}\epsilon_{a_1}}/A_t^2 & \cdots & \sigma_{\epsilon_{a_{J_t}}\epsilon_{a_{J_t}}}/A_t^2 & \sigma_{\epsilon_{a_{J_t}}\epsilon_{b_1}} & \cdots & \sigma_{\epsilon_{a_{J_t}}\epsilon_{b_{J_t}}} \\ \sigma_{\epsilon_{b_1}\epsilon_{a_1}} & \cdots & \sigma_{\epsilon_{b_1}\epsilon_{a_{J_t}}} & \sigma_{\epsilon_{b_1}\epsilon_{b_1}}A_t^2 & \cdots & \sigma_{\epsilon_{b_1}\epsilon_{b_{J_t}}}A_t^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\epsilon_{b_{J_t}}\epsilon_{a_1}} & \cdots & \sigma_{\epsilon_{b_{J_t}}\epsilon_{a_{J_t}}} & \sigma_{\epsilon_{b_{J_t}}\epsilon_{b_1}}A_t^2 & \cdots & \sigma_{\epsilon_{b_{J_t}}\epsilon_{b_J}}A_t^2 \end{bmatrix}_{2J_t \times 2J_t} .$$

Unfortunately, Eq. (3.12) cannot be directly calculated because $\mathbf{\Omega}$ is unknown. However, we can implement an iterative version of GLS similar to Feasible Generalized Least Squares (FGLS, Greene, 2003) to estimate β. A similar approach was used by Battauz and Bellio (2011), where the covariance matrix was computed on the basis of the information matrix for the estimated abilities. This procedure starts by using an initial estimate $\hat{\beta}$ to calculate $\hat{\mathbf{\Omega}}$, which can then be used to update $\hat{\beta}$ through GLS. This back-and-forth then continues until convergence of both $\hat{\beta}$ and $\hat{\mathbf{\Omega}}$.

In theory, any reasonable initial estimates for $\mathbf{a}^*$ and $\mathbf{b}^*$ can be used to kick-start the FGLS procedure. In practice, though, it is a good idea to either use commonly-observed values such as $a_j^* = 1$ and $b_j^* = 0 \ \forall j$ or even estimates given by a simpler equating procedure, such as those provided by the `equateMultiple` R package (Battauz, 2017b), which implements the methods from Battauz (2017a).

With those initial estimates, an initial $\mathbf{\Omega}$ can be computed; the explicit form for each of its blocks, $\mathbf{\Omega}_t$, can be found on Appendix B. Nevertheless, it has been our observation that the correlations between parameter estimates pertaining to different items are very close to zero. In these cases, a great deal of computational effort can be saved by estimating $\mathbf{\Omega}$ as per Thissen and Wainer (1982), which ignores the correlations between parameters from different items and performs calculations based on the joint maximum likelihood. This simplification comes to no observable penalty to the final estimation of β. Once an optimal solution for $(\mathbf{a}^*, \mathbf{b}^*)$ is obtained, the equating coefficients are estimated through maximization of the likelihood function in Eq. (3.11).

3.2.4 Calculating the standard errors of the equating coefficients

Since the point estimates of the equating coefficients are obtained by maximizing the likelihood function, we can apply standard asymptotic theory to estimate the standard errors (SE) of those coefficients (Bock and Aitkin, 1981; Bock and Lieberman, 1970). Based on that, Ogasawara (2001) worked out the explicit calculations the SEs of IRT equating coefficients using the delta method. This is also the method Battauz (2013, 2017a) used to obtain the covariance matrix of the equating coefficients and the synthetic item parameters.

Let $(\hat{\mathbf{A}}^T, \hat{\mathbf{B}}^T)$ be the vector containing the equating coefficient estimates and γ be the vector of the IRT-estimated parameters for all the items. Using the delta method, the asymptotic variance-covariance matrix for $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ is given by

$$\text{acov}\left[(\hat{\mathbf{A}}^T, \hat{\mathbf{B}}^T)^T\right] = \left[\frac{\partial(\mathbf{A}^T, \mathbf{B}^T)}{\partial\gamma}\right]^T \text{acov}(\hat{\gamma})\frac{\partial(\mathbf{A}^T, \mathbf{B}^T)}{\partial\gamma}.$$

The asymptotic covariance matrix of the item parameter estimates, acov$(\hat{\gamma})$, has dimension $2\sum_t J_t$; $\partial(\hat{\mathbf{A}}^T, \hat{\mathbf{B}}^T)/\partial\gamma$ has size $2\sum_t J_t \times 2T$.

Since the likelihood function does not provide an explicit form for the equating coefficients, these partial derivatives are not directly obtainable. However, they can be retrieved indirectly using the formula for partial derivatives in implicit functions,

$$\left[\frac{\partial(\mathbf{A}^T, \mathbf{B}^T)}{\partial\gamma}\right]^T = -\left[\frac{\partial S}{\partial(\mathbf{A}^T, \mathbf{B}^T)}\right]^{-1} \frac{\partial S}{\partial\gamma^T}. \tag{3.13}$$

Here, S is the gradient of the log-likelihood function, l_p, with respect to $(\mathbf{A}, \mathbf{B})$, i.e.:

$$S = \left[\frac{\partial l_p(\mathbf{A}, \mathbf{B})}{\partial(\mathbf{A}^T, \mathbf{B}^T)^T}\right] = 0.$$

Since S is the first derivative of l_p, its second derivative, $\partial S/\partial(\mathbf{A}^T, \mathbf{B}^T)$, will be the numerically-obtainable Hessian matrix

$$\frac{\partial S}{\partial(\mathbf{A}^T, \mathbf{B}^T)} = \frac{\partial^2 l_p(\mathbf{A}, \mathbf{B})}{\partial(\mathbf{A}^T, \mathbf{B}^T)^T \partial(\mathbf{A}^T, \mathbf{B}^T)}.$$

Likewise, the second term of Eq. (3.13), $\partial S/\partial\gamma^T$, will also correspond to a numerically-obtainable matrix of second derivatives. Specifically,

$$\frac{\partial S}{\partial\gamma^T} = \frac{\partial^2 l_p(\mathbf{A}, \mathbf{B})}{\partial(\mathbf{A}^T, \mathbf{B}^T)^T \partial\gamma^T}.$$

Table 3.1: Simulation scenarios. Those with multiple numbers of examinees (i.e., sc2 and sc4) perform a rotation in the number of examinees per form, with $I_1 = 2\,000$, $I_2 = 1\,000$, $I_3 = 500$, $I_4 = 2\,000$ again and so on.

Scenario	Test forms (T)	Items per form (J_t)	Examinees per form (I_t)
sc1	10	40	1 000
sc2	10	40	2 000, 1 000, 500
sc3	20	40	1 000
sc4	20	40	2 000, 1 000, 500

3.3 Simulation study and results

In order to study the properties of the proposed method, simulations were performed under different scenarios. The characteristics of each scenario are summarized on Tab. 3.1, and all software created and used to generate the data presented here are available upon request. The proposed method was developed in R (R Core Team, 2018) and C++ (Stroustrup, 2000) with the help of the following R packages, listed in no particular order: `mirt` (Chalmers, 2012), `equateIRT` (Battauz, 2015), `equateMultiple` (Battauz, 2017b), `Rcpp` (Eddelbuettel and Balamuta, 2017), `statmod` (Giner and Smyth, 2016), `mvtnorm` (Genz *et al.*, 2018; Genz and Bretz, 2009), `Matrix` (Bates and Maechler, 2018) and `numDeriv` (Gilbert and Varadhan, 2016), as well as `parallel` and `stats` (R Core Team, 2018).

All scenarios share the same distributional properties for the person and item parameters. The abilities were generated from Normal distributions, with the first form having $\theta \sim N(0, 1)$ and the last form having $\theta \sim N(0.5, 1.3^2)$. The parameters for the intermediate forms were picked from an equally-spaced vector of size T ranging from 0 and 0.5 for the mean and 1 and 1.3 for the standard deviations. As for the item parameters, we have discriminations uniformly-distributed in the $[0.5, 2]$ interval; the difficulty parameters come from a truncated standard Normal distribution, with $b \sim TN(0, 1, -2, 2)$.

The linkage plan was setup so that all forms would be linked to their four closest neighbors by having 5 items in common with them. The exceptions are the first and last two forms, which have less than four neighbors each. As an example, the linkage plan for sc1 is illustrated in Fig. 3.1, but the extension to the larger cases is straightforward. Likewise, Tab. 3.2 explicits the number of items in common between the test form in the row and the test form in the column. Again, extension to scenarios with more than 10 forms is straightforward.

The number of items per form was fixed at 40 because different numbers do not do much more than change the precision of the parameter estimates in all methods. In fact, during

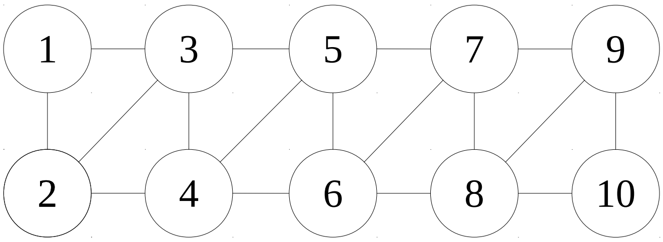


FIGURE 3.1: Linkage plan for scenario sc1. Each circle represents one test form; forms connected by a line have 5 items in common.

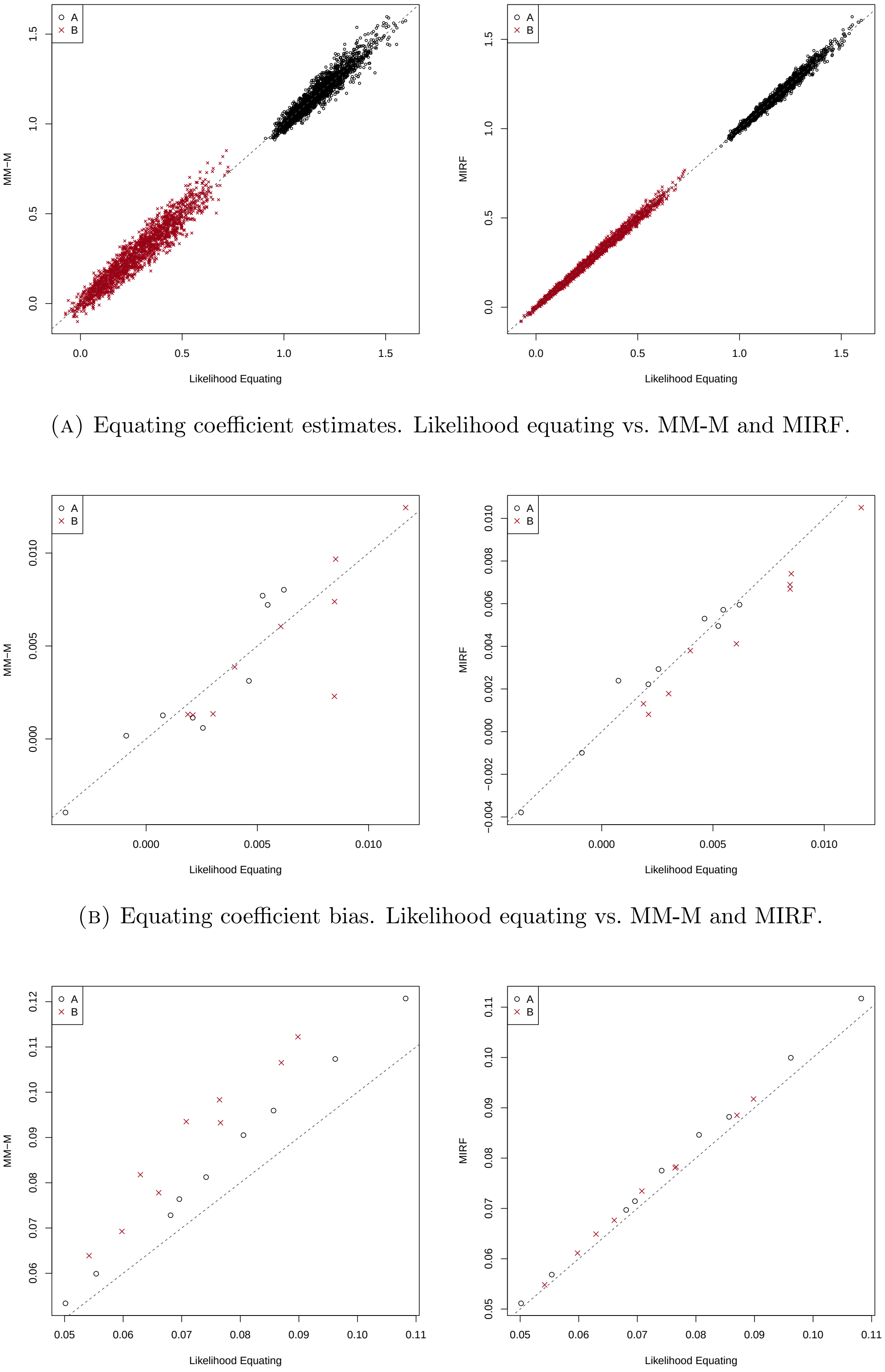
TABLE 3.2: Linkage plan for scenario sc1. Cells contain the quantity of items in common between the test form in the row and the one in the column.

	$t1$	$t2$	$t3$	$t4$	$t5$	$t6$	$t7$	$t8$	$t9$	$t10$
$t1$	40	5	5	-	-	-	-	-	-	-
$t2$	5	40	5	5	-	-	-	-	-	-
$t3$	5	5	40	5	5	-	-	-	-	-
$t4$	-	5	5	40	5	5	-	-	-	-
$t5$	-	-	5	5	40	5	5	-	-	-
$t6$	-	-	-	5	5	40	5	5	-	-
$t7$	-	-	-	-	5	5	40	5	5	-
$t8$	-	-	-	-	-	5	5	40	5	5
$t9$	-	-	-	-	-	-	5	5	40	5
$t10$	-	-	-	-	-	-	-	5	5	40

pretesting, using different values for J_t did not interfere with the conclusions regarding the comparison of the different equating methods.

Table 3.3 shows the bias and Root Mean Squared Error (RMSE) of the proposed method for each scenario. Figs. 3.2 to 3.5 compare the estimates, bias and RMSE of the proposed method with those observed by applying Multiple Mean-Mean (MM-M) and the Multiple Item Response Function (MIRF) from Battauz (2017a) to the same data. Each point in the top-row plots represents the estimate of one of the equating coefficients for one particular run of that scenario. Since $A_1 = 1$ and $B_1 = 0$ by definition, each plot will contain $(T - 1) \times 2 \times R$ points, where $R = 200$ runs of that scenario. Regarding the bias and RMSE of the equating coefficients, each point represents the average of A_t—or B_t—for a particular test form t across R runs of that scenario.

For the cases with variable I_t (sc2 and sc4), Fig. 3.6 presents scatterplots of the bias and RMSE for the estimates of $\mathbf{a}^*$ and $\mathbf{b}^*$. As before, the proposed method is laid horizontally against MM-M and MIRF. All unique items have been excluded. This figure might help highlight the differences in how each method calculates (a_j^*, b_j^*), which could accentuate the differences in the equating coefficients.

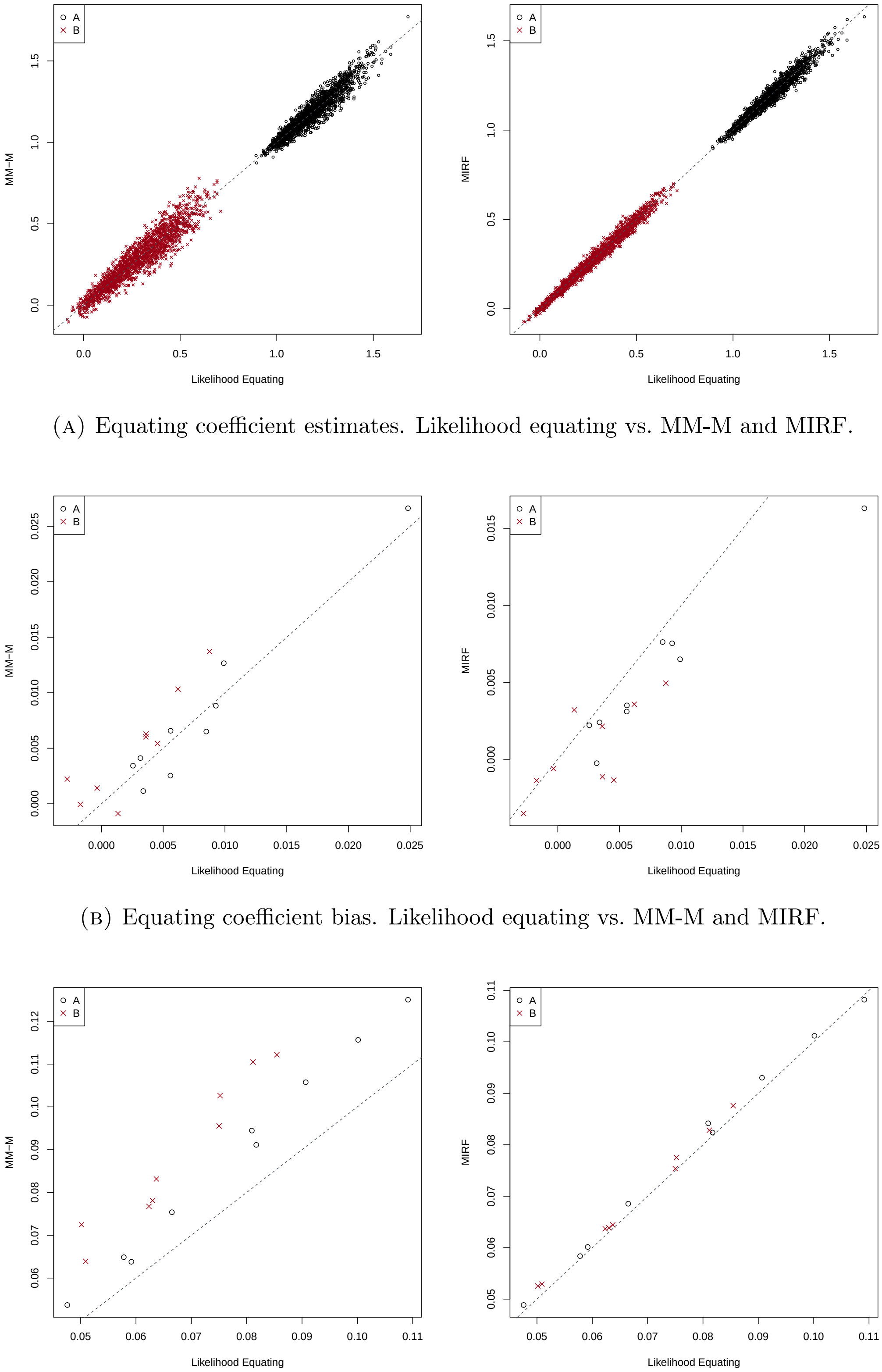


(A) Equating coefficient estimates. Likelihood equating vs. MM-M and MIRF.

(B) Equating coefficient bias. Likelihood equating vs. MM-M and MIRF.

(C) Equating coefficient RMSE. Likelihood equating vs. MM-M and MIRF.

FIGURE 3.2: Estimates, bias and RMSE of the equating coefficient estimates from likelihood equating against MM-M and MIRF. Scenario 1.

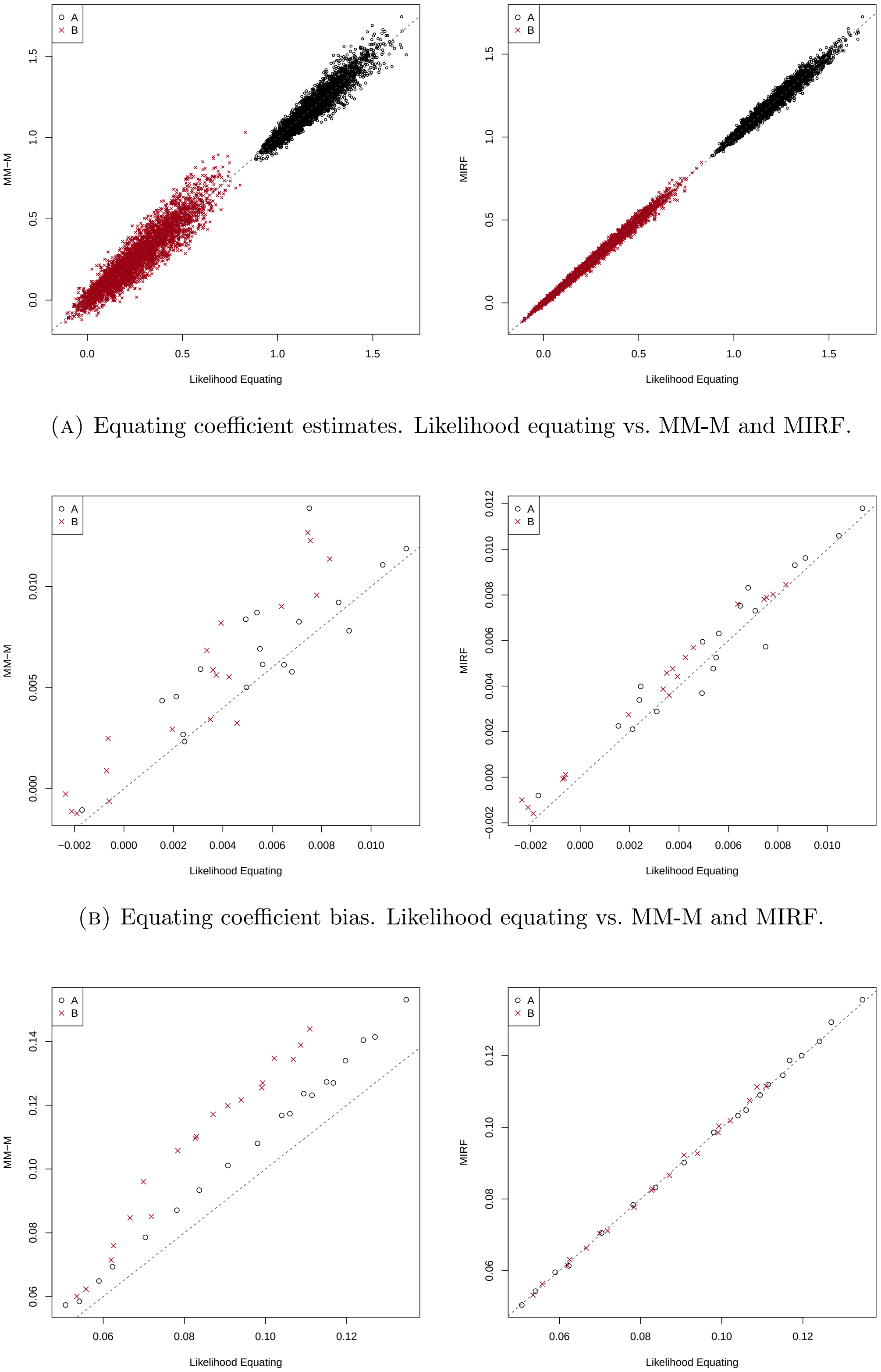


(A) Equating coefficient estimates. Likelihood equating vs. MM-M and MIRF.

(B) Equating coefficient bias. Likelihood equating vs. MM-M and MIRF.

(C) Equating coefficient RMSE. Likelihood equating vs. MM-M and MIRF.

FIGURE 3.3: Estimates, bias and RMSE of the equating coefficient estimates from likelihood equating against MM-M and MIRF. Scenario 2.

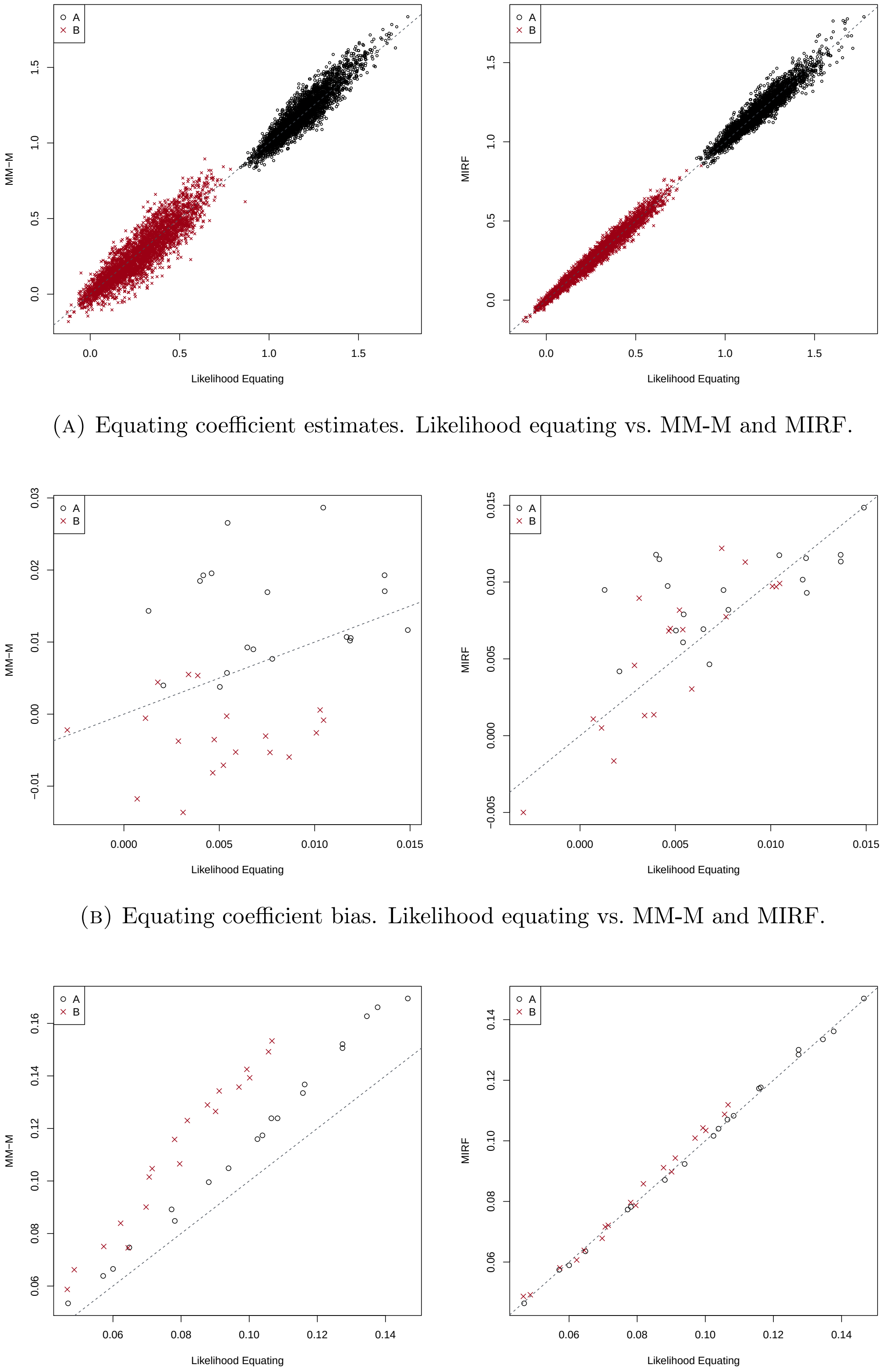


(A) Equating coefficient estimates. Likelihood equating vs. MM-M and MIRF.

(B) Equating coefficient bias. Likelihood equating vs. MM-M and MIRF.

(C) Equating coefficient RMSE. Likelihood equating vs. MM-M and MIRF.

Figure 3.4: Estimates, bias and RMSE of the equating coefficient estimates from likelihood equating against MM-M and MIRF. Scenario 3.

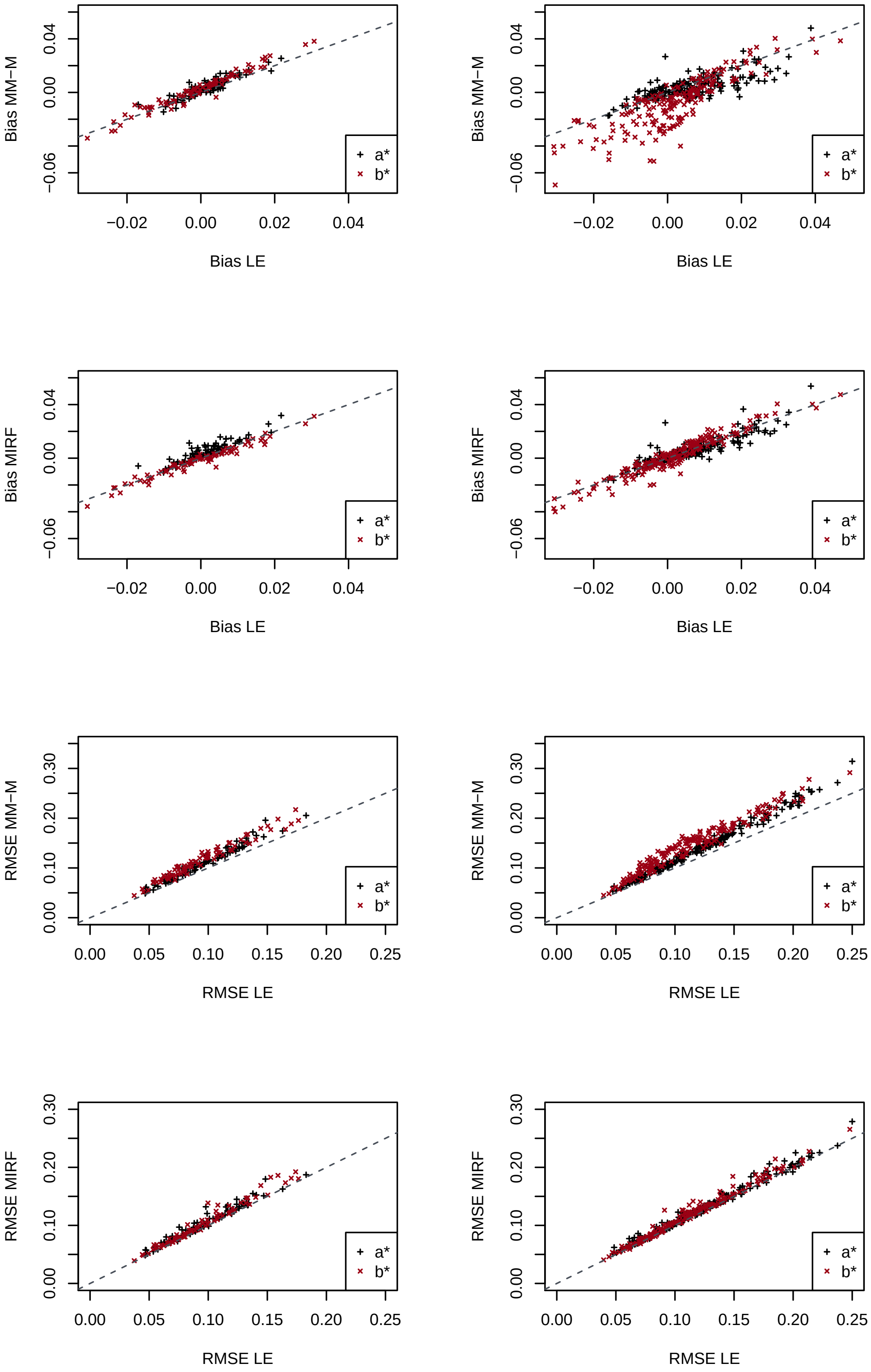


(A) Equating coefficient estimates. Likelihood equating vs. MM-M and MIRF.

(B) Equating coefficient bias. Likelihood equating vs. MM-M and MIRF.

(C) Equating coefficient RMSE. Likelihood equating vs. MM-M and MIRF.

FIGURE 3.5: Estimates, bias and RMSE of the equating coefficient estimates from likelihood equating against MM-M and MIRF. Scenario 4.



(A) Scenario 2.

(B) Scenario 4.

FIGURE 3.6: Bias and RMSE of $\hat{\mathbf{a}}^*$ and $\hat{\mathbf{b}}^*$ for common items. Likelihood equating (LE) against MM-M and MIRF. Scenarios 2 and 4.

Table 3.3: Bias and RMSE of likelihood equating per equating coefficient.

Scenario	Bias		RMSE	
	A	*B*	*A*	*B*
sc1	0.002489	0.006015	0.078315	0.072334
sc2	0.008083	0.002579	0.079615	0.068411
sc3	0.005466	0.003059	0.098780	0.085207
sc4	0.007821	0.004962	0.103501	0.081292
Overall	0.005965	0.004154	0.090053	0.076811

3.4 Discussion

The results from Tab. 3.3 suggest that the proposed method—likelihood equating, or LE for short—introduces a small amount of bias to the estimation of the equating coefficients. However, these values are arguably negligible, and were obtained on simulations of only 200 replications and samples of no more than 2 000 examinees, indicating good converging performance of LE.

The proposed method is also compared with two benchmark methods from Battauz (2017a), namely the Multiple Mean-Mean (MM-M) and the Multiple Item Response Function (MIRF) on Figs. 3.2 to 3.5 and 3.6. All scenarios show that the LE estimates have, in general, no more bias than MM-M or MIRF.

The RMSE plots on Figs. 3.2 to 3.5 are where the methodological innovation of LE begins to show its effects. Since LE accounts for some item properties that the state of the art does not, it is expected that RMSE plots favor LE over MM-M and MIRF. In effect, all scenarios showed that MM-M had higher RMSE for all equating coefficients. When compared to MIRF, the efficiency gain of LE is much less pronounced, perhaps even nonexistent in sc3. However, the gain in efficiency is quite small and does not increase with the number of test forms. Additionally, the equating coefficient estimates on subfigures (a) of the referred figures show that the equating coefficient estimates for LE and MIRF are much closer to each other than those for LE and MM-M, further reinforcing the similar performance between LE and MIRF.

Scenarios 1 and 3 respectively differ from 2 and 4 by how the even-numbered ones vary the number of examinees per form. This is done to introduce heteroskedasticity in the item parameter estimates. However, no discernible difference in method performance has been observed between them and the odd-numbered scenarios. These results defy the expectation that since both MM-M and MIRF assume homoskedasticity of item parameter estimates, scenarios where the number of examinees varies across the test forms should result in item parameter estimates with differing degrees of variance—i.e., heteroskedasticity—, which would make methods like LE perform noticeably better. Tests have been performed with scenarios containing a wider range of I_t, but those changes had little to no effect on the conclusions obtained from observing only the four scenarios published in this study.

Another explanation for the difference in behavior of LE when compared to MM-M and MIRF can be observed in Fig. 3.6, where the quality of the estimates of $\mathbf{a}^*$ and $\mathbf{b}^*$ for the common items is assessed. Once again, no discernible difference can be seen between the three methods regarding bias; when it comes to RMSE, though, LE performs distinctively better than MM-M, but also slightly better than MIRF. This can be attributed to the FGLS-like method implemented inside the LE procedure, which yielded more efficient synthetic parameter estimates than the closed-form solutions of the benchmark methods from Battauz (2017a), albeit at a computation cost. In the case of LE vs. MIRF, however, these already shy efficiency gains do not seem to translate into relevant differences in the equating coefficient estimates.

Equating parallel forms requires attention to aspects like differences in test difficulty or information function. We believe that the item parameter transformations through equating coefficients solves the problem of forms not being at the same difficulty level. The information function, however, is expected to have an impact on the standard errors.

In essence, LE provides a method for multiple-form equating based on a solid theoretical foundation. The statistical errors observed in the simulations are small and arguably innocuous, and LE can be a good alternative to state-of-the-art methods such as those proposed by Haberman (2009); Battauz (2013, 2017a), offering the potential of slightly higher efficiency without adding bias. Thus, this study succeeds in developing and implementing a new, statistically advantageous approach to multiple test equating. Nevertheless, the biggest practical cost of LE is its noticeably higher computational overhead. Using current, mid-tier technology, scenarios involving 50 forms with 100 items each often take hours to converge to an LE estimate for the equating coefficients, whereas methods like the MM-M and MIRF as implemented on `equateMultiple` usually take only a few seconds to perform their calculations. This is usually not a significant issue in practical applications, where the method only needs to be run once per administration, but it can become problematic in simulation scenarios, where methods are applied hundreds or even thousands of times over different samples.

All things considered, this novel foray into a multiple equating method that takes dependence and heteroskedasticity of item parameter estimates into consideration was shown to be insightful and rewarding. The scenarios studied contained the same number of total and common items, so future studies should study the sensitivity of the results to changes in those settings. New research efforts into the matter are also welcome, particularly investigations that (1) compare LE with other multiple equating methods under other simulation scenarios and real data applications or (2) attempt to further improve MML convergence speed. Aside from the velocity gains naturally obtained by the progress of computer hardware technology, speedier evolutions of LE can be developed by exploring alternative likelihood functions or better methods for estimating the synthetic parameters. The use of an iterative method (FGLS) to estimate parameters inside another iterative method (ML) can easily lead to computational bottlenecks, so perhaps the most promising immediate alternatives to LE can be discovered by applying different statistical methodologies, as opposed to attempting to achieve higher computational efficiency.

Appendix A

MLEs of a_j^* and b_j^*

Let us consider the likelihood function from Eq. (3.8). For one item j, we have the specific case

$$L = L\left(\mathbf{A}, \mathbf{B}, \mathbf{a}^*, \mathbf{b}^*, \sigma_a^2, \sigma_b^2; \hat{\mathbf{a}}, \hat{\mathbf{b}}\right) = \prod_{t=1}^{T} f(\hat{a}_{jt}) f(\hat{b}_{jt}), \qquad \text{(A.1)}$$

where $\mathbf{A}$ and $\mathbf{B}$ are vectors respectively containing the equating coefficients A_t and B_t for all $t = 1, \dots, T$ test forms. Likewise, $\mathbf{a}^* = \{a_1, \dots, a_j, \dots, a_J\}$, $\mathbf{b}^* = \{b_1, \dots, b_j, \dots, b_J\}$, $\hat{\mathbf{a}} = \{\hat{a}_{11}, \dots, \hat{a}_{jt}, \dots, \hat{a}_{JT}\}$ and $\hat{\mathbf{b}} = \{\hat{b}_{11}, \dots, \hat{b}_{jt}, \dots, \hat{b}_{JT}\}$.

Our goal is to find the maximum likelihood estimators of a_j^* and b_j^* in that function, i.e., the values of $\hat{a}_j^*$ and $\hat{b}_j^*$ that maximize the value of L.

Since $\hat{a}_{jt}$ and $\hat{b}_{jt}$ are independent from each other, finding the values of $\hat{a}_j^*$ and $\hat{b}_j^*$ that maximize L is a simple matter of finding the maximum of L with respect to each one of those parameters and then rewriting L as a function of them. To facilitate this derivation, we shall consider the one-to-one transformation below:

$$l = \log(L) = \log\left[\prod_{t=1}^{T} f(\hat{a}_{jt}) f(\hat{b}_{jt})\right] = \sum_{t=1}^{T} \log\left[f(\hat{a}_{jt})\right] + \sum_{t=1}^{T} \log\left[f(\hat{b}_{jt})\right] \qquad \text{(A.2)}$$

From Eq. (3.6), we know the distributional properties of $f(\hat{a}_{jt})$ and $f(\hat{b}_{jt})$. Specifically, we know that $\hat{a}_{jt} \overset{a}{\sim} N(A_t a_j^*, \sigma_a^2)$ and $\hat{b}_{jt} \overset{a}{\sim} N((b_j^* - B_t)/A_t, \sigma_b^2)$, where the variances were simplified due to the assumption of homoskedasticity. This means that

$$\log\left[f(\hat{a}_{jt})\right] = \log\left\{\frac{1}{\sqrt{2\pi}\sigma_a}\exp\left[\frac{-1}{2}\left(\frac{\hat{a}_{jt}-A_t a_j^*}{\sigma_a}\right)^2\right]\right\} \tag{A.3}$$

$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma_a}\right) - \frac{1}{2}\left(\frac{\hat{a}_{jt}-A_t a_j^*}{\sigma_a}\right)^2 \tag{A.4}$$

and

$$\log\left[f(\hat{b}_{jt})\right] = \log\left\{\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left[\frac{-1}{2}\left(\frac{\hat{b}_{jt}-(b_j^*-B_t)/A_t}{\sigma_b}\right)^2\right]\right\} \tag{A.5}$$

$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma_a}\right) - \frac{1}{2}\left(\frac{\hat{b}_{jt}A_t - b_j^* + B_t}{A_t\sigma_b}\right)^2. \tag{A.6}$$

The equations above tell us that l is composed by three different types of terms, but only one of those contains a_j^* and only one of them has b_j^*. This means that when deriving l with respect to those variables, we only have to deal with one simple sum in each case. Explicitly, we have that

$$\frac{\partial l}{\partial a_j^*} = \frac{\partial}{\partial a_j^*}\sum_{t=1}^{T}\left[-\frac{1}{2}\left(\frac{\hat{a}_{jt}-A_t a_j^*}{\sigma_a}\right)^2\right] \tag{A.7}$$

$$= \sum_{t=1}^{T}\left[-\frac{2}{2}\left(\frac{\hat{a}_{jt}-A_t a_j^*}{\sigma_a}\right)\left(\frac{-A_t}{\sigma_a}\right)\right] \tag{A.8}$$

$$= \sum_{t=1}^{T}\left(\frac{\hat{a}_{jt}A_t - A_t^2 a_j^*}{\sigma_a}\right), \tag{A.9}$$

which gives us the MLE of a_j^* from the following operation:

$$\frac{\partial l}{\partial a_j^*} = 0 \tag{A.10}$$

$$\sum_{t=1}^{T} \frac{\hat{a}_{jt} A_t - A_t^2 \hat{a}_j^*}{\sigma_a} = 0 \tag{A.11}$$

$$\sum_{t=1}^{T} \left(\hat{a}_{jt} A_t - A_t^2 \hat{a}_j^* \right) = 0 \tag{A.12}$$

$$\sum_{t=1}^{T} \hat{a}_{jt} A_t - \sum_{t=1}^{T} A_t^2 \hat{a}_j^* = 0 \tag{A.13}$$

$$\sum_{t=1}^{T} \hat{a}_{jt} A_t = \sum_{t=1}^{T} A_t^2 \hat{a}_j^* \tag{A.14}$$

$$\sum_{t=1}^{T} \hat{a}_{jt} A_t = \hat{a}_j^* \sum_{t=1}^{T} A_t^2 \tag{A.15}$$

$$\frac{\sum_{t=1}^{T} \hat{a}_{jt} A_t}{\sum_{t=1}^{T} A_t^2} = \hat{a}_j^* \tag{A.16}$$

For b_j^*, we have

$$\frac{\partial l}{\partial b_j^*} = \frac{\partial}{\partial b_j^*} \sum_{t=1}^{T} \left[-\frac{1}{2} \left(\frac{\hat{b}_{jt} A_t - b_j^* + B_t}{A_t \sigma_b} \right)^2 \right] \tag{A.17}$$

$$= \sum_{t=1}^{T} \left[-\frac{2}{2} \left(\frac{\hat{b}_{jt} A_t - b_j^* + B_t}{A_t \sigma_b} \right) \left(\frac{-1}{A_t \sigma_b} \right) \right] \tag{A.18}$$

$$= \sum_{t=1}^{T} \left(\frac{\hat{b}_{jt} A_t - b_j^* + B_t}{A_t^2 \sigma_b^2} \right), \tag{A.19}$$

which allows us to calculate the MLE $\hat{b}_j^*$ as follows:

$$\frac{\partial l}{\partial b_j^*} = 0 \tag{A.20}$$

$$\sum_{t=1}^{T} \left(\frac{\hat{b}_{jt} A_t - \hat{b}_j^* + B_t}{A_t^2 \sigma_b^2} \right) = 0 \tag{A.21}$$

$$\sum_{t=1}^{T} \left(\frac{\hat{b}_{jt} A_t - \hat{b}_j^* + B_t}{A_t^2} \right) = 0 \tag{A.22}$$

$$\sum_{t=1}^{T} \frac{\hat{b}_{jt} A_t + B_t}{A_t^2} - \sum_{t=1}^{T} \frac{\hat{b}_j^*}{A_t^2} = 0 \tag{A.23}$$

$$\sum_{t=1}^{T} \frac{\hat{b}_{jt} A_t + B_t}{A_t^2} = \sum_{t=1}^{T} \frac{\hat{b}_j^*}{A_t^2} \tag{A.24}$$

$$\sum_{t=1}^{T} \frac{\hat{b}_{jt} A_t + B_t}{A_t^2} = \hat{b}_j^* \sum_{t=1}^{T} \frac{1}{A_t^2} \tag{A.25}$$

$$\frac{\sum_{t=1}^{T} \frac{\hat{b}_{jt} A_t + B_t}{A_t^2}}{\sum_{t=1}^{T} \frac{1}{A_t^2}} = \hat{b}_j^* \tag{A.26}$$

After some straightforward rearrangements, equations (A.16) and (A.26) give us

$$\hat{a}_j^* = \frac{\sum_t^T \frac{\hat{a}_{jt}}{A_t} A_t^2}{\sum_t A_t^2} \tag{A.27}$$

and

$$\hat{b}_j^* = \frac{\sum_t^T \frac{\hat{b}_{jt} A_t + B_t}{A_t^2}}{\sum_t^T \frac{1}{A_t^2}}. \tag{A.28}$$

Appendix B

Covariance matrix for the item parameter estimates

B.1 Introduction

Let the function $L(\mathbf{a}, \mathbf{b}; \mathbf{x})$ determine the likelihood of the item and person parameter estimates given the test answers. For the J_t items in a particular form t, $\mathbf{a} = \{a_1, \ldots, a_{J_t}\}$ and $\mathbf{b} = \{b_1, \ldots, b_{J_t}\}$ are the vectors of the J_t item parameters and $\mathbf{x}$ is an $I \times J_t$ matrix containing the dichotomous item answers for I test takers. For the sake of readability, let $J = J_t$ for the rest of this section. The functional form of L in this context is

$$L = L(\mathbf{a}, \mathbf{b}; \mathbf{x}) = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta_i, \tag{B.1}$$

where P_{ij} is equivalent to $\Pr(X_{ijt} = 1 | \theta_{it}; a_{jt}, b_{jt})$ from Eq.(3.1) with t fixed and thus dropped out of the equation, θ being integrated out and a, b and x being vectorized. Moreover, $Q_{ij} = 1 - P_{ij}$ and, since θ is assumed to be normally-distributed, $f(\theta_i)$ is known and well-defined.

Let $\mathbf{\Sigma}$ be the covariance matrix of the IRT item parameter estimations $\hat{a}_{jt}$ and $\hat{b}_{jt}$. According to Kendall and Stuart (1966), the asymptotic $\mathbf{\Sigma}$ is given by the inverse of the negative expected value of the Hessian matrix of the log-likelihood function.

Let l be the logarithm of the likelihood function from Eq. (B.1). Moreover, let $\mathbf{H}$ be the Hessian matrix

$$\mathbf{H} = \begin{bmatrix} \partial^2 l/\partial a_1 \partial a_1 & \cdots & \partial^2 l/\partial a_1 \partial a_J & \partial^2 l/\partial a_1 \partial b_1 & \cdots & \partial^2 l/\partial a_1 \partial b_J \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \partial^2 l/\partial a_J \partial a_1 & \cdots & \partial^2 l/\partial a_J \partial a_J & \partial^2 l/\partial a_J \partial b_1 & \cdots & \partial^2 l/\partial a_J \partial b_J \\ \partial^2 l/\partial b_1 \partial a_1 & \cdots & \partial^2 l/\partial b_1 \partial a_J & \partial^2 l/\partial b_1 \partial b_1 & \cdots & \partial^2 l/\partial b_1 \partial b_J \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \partial^2 l/\partial b_J \partial a_1 & \cdots & \partial^2 l/\partial b_J \partial a_J & \partial^2 l/\partial b_J \partial b_1 & \cdots & \partial^2 l/\partial b_J \partial b_J \end{bmatrix}. \tag{B.2}$$

Once all the second derivatives and their expected values have been calculated, the covariance matrix of the item parameter estimates is assembled as

$$\mathbf{\Sigma} = -E(\mathbf{H}). \tag{B.3}$$

The individual elements of $\mathbf{H}$ are calculated in the following sections.

B.2 First derivatives

Let the item parameters a_k and b_k for some $k = 1, \ldots, J$ be represented by a generic parameter ξ_k. The first derivative of l with respect to ξ_k is given by

$$\frac{\partial l}{\partial \xi_k} = \frac{\partial}{\partial \xi_k} \sum_{i=1}^{I} \log \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta \tag{B.4}$$

$$= \sum_{i=1}^{I} \frac{\partial}{\partial \xi_k} \log \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta. \tag{B.5}$$

Let us name the integral above, $\int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta$, as a certain function g. Then we have:

$$\frac{\partial l}{\partial \xi_k} = \sum_{i=1}^{I} \frac{\partial}{\partial \xi_k} \log \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta \tag{B.6}$$

$$= \sum_{i=1}^{I} \frac{\partial}{\partial \xi_k} \log g. \tag{B.7}$$

$$= \sum_{i=1}^{I} \frac{1}{g} \frac{\partial g}{\partial \xi_k}. \tag{B.8}$$

Developing $\partial g/\partial \xi_k$, considering $\int \equiv \int_{-\infty}^{\infty}$ from now on for convenience:

$$\frac{\partial g}{\partial \xi_k} = \frac{\partial}{\partial \xi_k} \int \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta \tag{B.9}$$

$$= \int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left(\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} \right) f(\theta_i) d\theta, \tag{B.10}$$

It is useful to separately define $\partial P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} / \partial \xi_k$, which the product rule of differentiation allows us to equate it to

$$x_{ik} \left(\frac{P_{ik}}{Q_{ik}} \right)^{x_{ik}-1} \frac{\partial P_{ik}}{\partial \xi_k} - (1 - x_{ik}) \left(\frac{P_{ik}}{Q_{ik}} \right)^{x_{ik}} \frac{\partial P_{ik}}{\partial \xi_k}. \tag{B.11}$$

Remembering that $x_{ik} = \{0, 1\}$ for all $i = 1, \ldots, I$ and $k = 1, \ldots, J$, we can greatly simplify the equation above. For $x_{ik} = 1$:

$$\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} = \left[1 \left(\frac{P_{ik}}{Q_{ik}} \right)^0 \frac{\partial P_{ik}}{\partial \xi_k} - 0 \left(\frac{P_{ik}}{Q_{ik}} \right)^1 \frac{\partial P_{ik}}{\partial \xi_k} \right] = \frac{\partial P_{ik}}{\partial \xi_k}. \tag{B.12}$$

Likewise, $x_{ik} = 0$ gives us

$$\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} = \left[0 \left(\frac{P_{ik}}{Q_{ik}} \right)^1 \frac{\partial P_{ik}}{\partial \xi_k} - 1 \left(\frac{P_{ik}}{Q_{ik}} \right)^0 \frac{\partial P_{ik}}{\partial \xi_k} \right] = -\frac{\partial P_{ik}}{\partial \xi_k}. \tag{B.13}$$

Joining equations (B.12) and (B.13) with the signaling term $(2x_{ik} - 1)$, we have:

$$\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} = (2x_{ik} - 1) \frac{\partial P_{ik}}{\partial \xi_k}. \tag{B.14}$$

Finally, we put equations (B.8), (B.10) and (B.14) together to obtain the first derivatives of the likelihood function in Eq. (B.1) with respect to the generic item parameter ξ_k:

$$\frac{\partial l}{\partial \xi_k} = \sum_{i=1}^{I} \frac{\int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ik} - 1) \frac{\partial P_{ik}}{\partial \xi_k} \right] f(\theta_i) d\theta}{\int \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta}. \tag{B.15}$$

The functional form of the partial derivative $\partial P_{ik}/\partial \xi_k$ seen above depends on whether ξ_k corresponds to the discrimination or the difficulty parameter. They were derived in

Thissen and Wainer (1982), but can be verified without much effort.

$$\xi_k \equiv a_k \implies \frac{\partial P_{ik}}{\partial a_k} = P_{ik}Q_{ik}(\theta_i - b_k) \tag{B.16}$$

$$\xi_k \equiv b_k \implies \frac{\partial P_{ik}}{\partial a_k} = P_{ik}Q_{ik}(-a_k). \tag{B.17}$$

B.3 Second derivatives

Let η_h be a second representation of a generic parameter of item h; h can be equal or different from k from the previous section. This allows us to generalize all the second derivatives composing the Hessian matrix on Eq. (B.2) as $\partial l/(\partial\xi_k\partial\eta_h)$, remembering that this is equivalent to $\partial l/(\partial\eta_h\partial\xi_k)$. Since Eq. (B.8) gives $\partial l/\partial\xi_k = 1/g \cdot (\partial g/\partial\xi_k)$,

$$\frac{\partial^2 l}{\partial\xi_k\partial\eta_h} = \frac{\partial l}{\partial\eta_h}\frac{\partial l}{\partial\xi_k} = \frac{\partial l}{\partial\eta_h}\sum_{i=1}^{I}\frac{1}{g}\frac{\partial g}{\partial\xi_k} = \sum_{i=1}^{I}\frac{1}{g^2}\left(g\frac{\partial^2 g}{\partial\eta_h\partial\xi_k} - \frac{\partial g}{\partial\xi_k}\frac{\partial g}{\partial\eta_h}\right). \tag{B.18}$$

It should be noted that where $\partial g/\xi_k$ and g are respectively the numerator and denominator from Eq. (B.15). Moreover, $\partial g/\partial\eta_h$ has already been defined in Eq. (B.10), and it works for η_h equal to ξ_k or different from it because both are actually just placeholders for the item parameters a_j and b_j. Therefore, the only piece left for development on Eq. (B.18) is $\partial^2 g/(\partial\eta_h\partial\xi_k)$. We begin by expanding it:

$$\frac{\partial^2 g}{\partial\eta_h\partial\xi_k} = \frac{\partial}{\partial\eta_h}\int\left(\prod_{j\neq k}P_{ij}^{x_{ij}}Q_{ij}^{1-x_{ij}}\right)\left(\frac{\partial}{\partial\xi_k}P_{ik}^{x_{ik}}Q_{ik}^{1-x_{ik}}\right)f(\theta_i)d\theta. \tag{B.19}$$

Further expansion of the equation above depends on whether h is equal to k or not. We will begin with the arguably simpler case of $h \neq k$:

$$\frac{\partial^2 g}{\partial\eta_h\partial\xi_k} = \frac{\partial}{\partial\eta_h}\int\left(\prod_{j\neq k}P_{ij}^{x_{ij}}Q_{ij}^{1-x_{ij}}\right)\left(\frac{\partial}{\partial\xi_k}P_{ik}^{x_{ik}}Q_{ik}^{1-x_{ik}}\right)f(\theta_i)d\theta = \tag{B.20}$$

$$\int\frac{\partial}{\partial\eta_h}\left(\prod_{j\neq k}P_{ij}^{x_{ij}}Q_{ij}^{1-x_{ij}}\right)\left(\frac{\partial}{\partial\xi_k}P_{ik}^{x_{ik}}Q_{ik}^{1-x_{ik}}\right)f(\theta_i)d\theta = \tag{B.21}$$

$$\int\frac{\partial}{\partial\eta_h}\left(\prod_{j\neq k}P_{ij}^{x_{ij}}Q_{ij}^{1-x_{ij}}\right)\left[(2x_{ik}-1)\frac{\partial P_{ik}}{\partial\xi_k}\right]f(\theta_i)d\theta = \tag{B.22}$$

$$\int\left(\prod_{j\neq k\neq h}P_{ij}^{x_{ij}}Q_{ij}^{1-x_{ij}}\right)\left[(2x_{ik}-1)\frac{\partial P_{ik}}{\partial\xi_k}\right]\left[(2x_{ih}-1)\frac{\partial P_{ih}}{\partial\eta_h}\right]f(\theta_i)d\theta, \tag{B.23}$$

For the case of $h = k$, we have $\eta_h = \eta_k$, which does not necessarily imply that $\eta_k = \xi_k$ because each one can be referring to a different parameter of the same item k.

$$\frac{\partial^2 g}{\partial \eta_h \partial \xi_k} = \frac{\partial}{\partial \eta_k} \int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left(\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} \right) f(\theta_i) d\theta = \tag{B.24}$$

$$\int \frac{\partial}{\partial \eta_k} \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left(\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} \right) f(\theta_i) d\theta = \tag{B.25}$$

$$\int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \frac{\partial}{\partial \eta_k} \left(\frac{\partial}{\partial \xi_k} P_{ik}^{x_{ik}} Q_{ik}^{1-x_{ik}} \right) f(\theta_i) d\theta = \tag{B.26}$$

$$\int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \frac{\partial}{\partial \eta_k} \left[(2x_{ih} - 1) \frac{\partial P_{ih}}{\partial \eta_k} \right] = \tag{B.27}$$

$$\int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ik} - 1) \frac{\partial^2 P_{ik}}{\partial \xi_k \partial \eta_k} \right] f(\theta_i) d\theta \tag{B.28}$$

From Eq. (B.16), we know that $\partial^2 P_{ik} / (\partial \xi_k \partial \eta_k)$ will be equal to zero whenever ξ_k and η_h are referring to different item parameters, i.e., one is the discrimination and the other is the difficulty. Moreover, that equation shows that if they refer to the same item parameter, then

$$\frac{\partial^2 P_{ik}}{\partial \xi_k \partial \eta_k} = -P_{ik} Q_{ik}. \tag{B.29}$$

Now we can finally assemble the second derivatives of the log-likelihood function. Equations (B.18), (B.23) and (B.28) give us

$$\frac{\partial^2 l}{\partial \xi_k \partial \eta_h} = \sum_{i=1}^{I} \frac{1}{g^2} \left(g \frac{\partial^2 g}{\partial \eta_h \partial \xi_k} - \frac{\partial g}{\partial \xi_k} \frac{\partial g}{\partial \eta_h} \right), \tag{B.30}$$

where

$$g = \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} f(\theta_i) d\theta \tag{B.31}$$

$$\frac{\partial g}{\xi_k} = \int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ik} - 1) \frac{\partial P_{ik}}{\partial \xi_k} \right] f(\theta_i) d\theta \tag{B.32}$$

$$\frac{\partial^2 g}{\partial \eta_h \partial \xi_k} = \begin{cases} \int \left(\prod_{j \neq k} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ik} - 1) \frac{\partial^2 P_{ik}}{\partial \xi_k \partial \eta_k} \right] f(\theta_i) d\theta & h = k \\ \int \left(\prod_{j \neq k \neq h} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ik} - 1) \frac{\partial P_{ik}}{\partial \xi_k} \right] & \\ \qquad \left[(2x_{ih} - 1) \frac{\partial P_{ih}}{\partial \eta_h} \right] f(\theta_i) d\theta & h \neq k \end{cases} \tag{B.33}$$

$$\frac{\partial g}{\partial \eta_h} = \int \left(\prod_{j \neq h} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \right) \left[(2x_{ih} - 1) \frac{\partial P_{ih}}{\partial \eta_h} \right] f(\theta_i) d\theta \tag{B.34}$$

B.4 Expected values

In their work, Thissen and Wainer (1982) worked with a log-likelihood function defined as

$$\sum_{j=1}^{J} \sum_{i=1}^{I} x_{ij} \log(P_{ij}) + (1 - x_{ij}) \log(Q_{ij}). \tag{B.35}$$

This is a linear function of the item response x_{ij}, which makes the calculation of the expected value of the Hessian matrix straightforward. However, when we marginalize θ, the result is a non-linear log-likelihood function. Nonetheless, $E(\mathbf{H})$ can still be calculated.

Let $(x_{i1}, \ldots, x_{iJ})$ be the set of binary responses given by person i in a test. Also, let $m(x_{i1}, \ldots, x_{iJ})$ be the second derivative of the log-likelihood function with respect to the item parameter estimates for that person i, i.e.,

$$m(x_{i1}, \ldots, x_{iJ}) = \frac{\partial^2 l}{\partial \xi_k \partial \eta_h}. \tag{B.36}$$

Then, the expected value of $\mathbf{H}$ can be defined as

$$\begin{aligned} E(\mathbf{H}) &= nE[m(x_{i1}, \ldots, x_{iJ})] & \text{(B.37)} \\ &= n \sum_{x_{i1}} \cdots \sum_{x_{iJ}} \left[m(x_{i1}, \ldots, x_{iJ}) \Pr(x_{i1}, \ldots, x_{iJ}) \right]. & \text{(B.38)} \end{aligned}$$

Bibliography

Aitkin, A. (1935) On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55**, 42–48.

Andersson, B., Bränberg, K. and Wiberg, M. (2013) Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software* **55**(6), 1–25.

Andersson, B., Bränberg, K. and Wiberg, M. (2014) *Test equating using the kernel method with the R package kequate.* R: Vignette.

Andersson, B. and Wiberg, M. (2017) Item response theory observed-score kernel equating. *Psychometrika* **82**(1), 48–66.

Angoff, W. (1971) Scales, norms, and equivalent scores. in rl thorndike (ed.), educational measurement . .

Bartolucci, F., Bacci, S. and Gnaldi, M. (2016) *Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata.* ISBN 978-1-4665-6849-5.

Bates, D. and Maechler, M. (2018) *Matrix: Sparse and Dense Matrix Classes and Methods.* R package version 1.2-14.

Battauz, M. (2013) IRT Test Equating in Complex Linkage Plans. *Psychometrika* **78**(3), 464–480.

Battauz, M. (2015) equateIRT: An R package for IRT test equating. *Journal of Statistical Software* **68**(7), 1–22.

Battauz, M. (2017a) Multiple equating of separate IRT calibrations. *Psychometrika* **82**(3), 610–636.

Battauz, M. (2017b) *equateMultiple: Equating of Multiple Forms.* R package version 0.0.0.

Battauz, M. and Bellio, R. (2011) Structural modeling of measurement error in generalized linear models with Rasch measures as covariates. *Psychometrika* **76**(1), 40–56.

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

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

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

Braun, H. I. and Holland, P. W. (1982) Observed-score test equating: a mathematical analysis of some ets equating procedures. In *Test equating*, eds P. W. Holland and D. B. Rubin, volume 1, pp. 9–49. New York: Academic Press.

Chalmers, R. P. (2012) mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* **48**(6), 1–29.

Crocker, L. and Algina, J. (2008) *Introduction to Classical and Modern Test Theory*. Cengage Learning.

von Davier, A. A. (2010) Equating observed-scores: The percentile rank, gaussian kernel, and IRT observed-score equating methods. In *International Meeting of Psychometric Society*.

von Davier, A. A. and Chen, H. (2013) The Kernel Levine Equipercentile Observed-Score Equating Function. Technical report, Educational Testing Service, Princeton, New Jersey.

von Davier, A. A., Holland, P. W. and Thayer, D. T. (2004) *The Kernel Method of Test Equating*. New York: Springer. ISBN 0387019855.

Dorans, N. J. and Feigenbaum, M. D. (1994) Equating issues engendered by changes to the SAT and PSAT/NMSQT. *Technical Issues Related to the Introduction of the New SAT and PSAT/NMSQT* pp. 91–122.

Eddelbuettel, D. and Balamuta, J. J. (2017) Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints* **5**, e3188v1.

Genz, A. and Bretz, F. (2009) *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag. ISBN 978-3-642-01688-2.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2018) *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-8.

Gilbert, P. and Varadhan, R. (2016) *numDeriv: Accurate Numerical Derivatives.* R package version 2016.8-1.

Giner, G. and Smyth, G. K. (2016) statmod: probability calculations for the inverse gaussian distribution. *R Journal* **8**(1), 339–351.

González, J. and von Davier, M. (2013) Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement* **50**(3), 315–320.

González, J. and Wiberg, M. (2017) *Applying Test Equating Methods Using R.* New York: Springer.

González, J., Wiberg, M. and von Davier, A. A. (2016) A note on the Poisson's binomial distribution in item response theory. *Applied Psychological Measurement* **40**(4), 302–310.

Greene, W. H. (2003) *Econometric Analysis.* Pearson Education India.

Gulliksen, H. (1950) Wiley publications in psychology. Theory of mental tests.

Haberman, S. J. (2009) Linking Parameter Estimates Derived From an Item-Response Model Through Separate Calibrations. Technical Report December, Educational Testing Service.

Haebara, T. (1980) Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research* **22**(3), 144–149.

Hambleton, R. K. and Jones, R. W. (1993) Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice* **12**(3), 38–47.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991) *Fundamentals of Item Response Theory.* Volume 2. Sage.

Han, T., Kolen, M. and Pohlmann, J. (1997) A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education* **10**(2), 105–121.

Hanson, B. A. (1996) Testing for Differences in Test Score Distributions Using Loglinear Models. *Applied Measurement in Education* **9**(4), 305–321.

Harris, D. J. and Crouse, J. D. (1993) A study of criteria used in equating. *Applied Measurement in Education* **6**(3), 195–240.

Holland, P. W. and Thayer, D. T. (1987) Notes on the use of log-linear models for fitting discrete probability distributions. *ETS Research Report Series* **1987**(2), i–40.

Holland, P. W. and Thayer, D. T. (2000) Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics* **25**(2), 133–183.

Jiang, Y., von Davier, A. A. and Chen, H. (2012) Evaluating equating results: percent relative error for chained kernel equating. *Journal of Educational Measurement* **49(1)**, 39–58.

Johnson, R. and Kuby, P. (2008) *Elementary Statistics.* 10th edition. Berlmont, CA, USA: Thomsom Brooks/Cole. ISBN 978-0-495-38386-4.

Kendall, M. G. and Stuart, A. (1966) The advanced theory of statistics. *Inference and Relationship, Charles Griffin & Company Ltd., London* **2**.

Kim, S.-H. and Cohen, A. S. (1998) A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement* **22**(2), 131–143.

Kolen, M. J. and Brennan, R. L. (2014) *Test Equating, Scaling, and linking: Methods and Practices.* Third edition. New York: Springer.

Levine, R. (1955) Equating the score scales of alternate forms administered to samples of different ability. *ETS Research Bulletin Series* **1955**(2), i–118.

Leôncio, W. and Wiberg, M. (2018) Evaluating Equating Transformations from Different Frameworks. In *Springer Proceedings in Mathematics & Statistics*, volume 233.

van der Linden, W. J. (2011) Local observed-score equating. In *Statistical models for test equating, scaling, and linking*, ed. A. von Davier, pp. 201–223. New York: Springer.

van der Linden, W. J. and Barrett, M. D. (2016) Linking Item Response Model Parameters. *Psychometrika* **81**(3), 650–673.

van der Linden, W. J. and Hambleton, R. K. (1998) *Handbook of Modem Item Response Theory.* ISBN 9781441928498.

Lord, F. M. (1953) The relation of test score to the trait underlying the test. *Educational and Psychological Measurement* **13**(4), 517–549.

Lord, F. M. (1977) Practical applications of item characteristic curve theory. *Journal of Educational Measurement* **14**(2), 177–138.

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

Lord, F. M. and Novick, M. R. (1967) *Statistical Theories of Mental Test Scores.* ISBN 9781593119348.

Lord, F. M. and Wingersky, M. S. (1984) Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement* **8**(4), 453–461.

Loyd, B. H. and Hoover, H. (1980) Vertical equating using the Rasch model. *Journal of Educational Measurement* **17**(3), 179–193.

Marco, G. L. (1977) Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement* **14**(2), 139–160.

Meng, Y. (2012) *Comparison of kernel equating and item response theory equating methods.* Dissertation submitted to the graduate school of the university of massachusetts amherst in partial fulfillment of the requirements for the degree of doctor of education, University of Massachusetts Amherst.

Ogasawara, H. (2001) Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement* **25**(1), 53–67.

Ogasawara, H. (2003) Asymptotic standard errors of IRT observed-score equating methods. *Society* **68**(2), 193–211.

R Core Team (2018) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reise, S. P. and Revicki, D. A. (2015) *Handbook of Item Response Theory Modeling: Applications to typical performance assessment.* ISBN 9781848729728.

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

Rosenbaum, P. R. and Thayer, D. (1987) Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology* **40**(1), 43–49.

Stocking, M. L. and Lord, F. M. (1982) Developing a common metric in item response theory. *ETS Research Report Series* **1982**(1).

Stocking, M. L. and Lord, F. M. (1983) Developing a common metric in item response theory. *Applied Psychological Measurement* **7**(2), 201–210.

Stroustrup, B. (2000) *The C++ Programming Language.* Pearson Education India.

Thissen, D. and Wainer, H. (1982) Some standard errors in item response theory. *Psychometrika* **47**(4), 397–412.

Wang, T., Hanson, B. A. and Harris, D. J. (2000) The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement* **24**(3), 195–210.

Wiberg, M. (2016) Alternative linear item response theory observed-score equating methods. *Applied Psychological Measurement* **40**(3), 180–199.

Wiberg, M. and González, J. (2016) Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement* **53**(1), 106–125.

Wiberg, M., van der Linden, W. J. and von Davier, A. A. (2014) Local observed-score kernel equating. *Journal of Educational Measurement* **51**, 57–74.

Waldir Leôncio Netto

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +47 412 09 531
e-mail: waldir.leoncionetto@phd.unipd.it

Current Position

Since October 2015 (expected completion: 3/2019):
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Advances in test equating: comparing IRT and Kernel methods and a new likelihood approach to equate multiple forms
Supervisor: Prof. Michela Battauz (Università Degli Studi di Udine)
Co-supervisor: Prof. Marie Wiberg (Umeå Universitet).

Research interests

- Computational Statistics
- Item Response Theory
- Automatic Item Generation
- Item pre-calibration
- Test equating

Education

2001 – 2005
Bachelor degree in Statistics.
University of Brasília, Department of Statistics
Title of dissertation: "Prediction models in time series: an application to the Ibovespa"
Supervisor: Prof. Geraldo da Silva e Souza.

Visiting periods

01/2017 – 06/2017
Department of Statistics, Umeå University,
Umeå, Sweden.
Supervisor: Prof. Marie Wiberg

01/2018 – 09/2018
Centre for Educational Measurement, University of Oslo,
Oslo, Norway.
Supervisor: Prof. Björn Andersson

Work experience

01/2008 – 09/2015
Oficina das Finanças.
Personal Financial Planner

08/2007 – 09/2015
Attorney-General's Office (Brazil).
Statistician

12/2005 – 01/2007
Department of Public Safety (Brasília, Brazil).
Administrative Assistant

Computer skills

- Programming languages: R, C++, Python, VBA
- Document preparation: LaTeX, LibreOffice, Microsoft Office
- Operating systems: Linux, Windows, macOS

Language skills

Brazilian Portuguese (native), English (fluent), Italian (fluent), German (basic), Norwegian (basic)

Publications

Articles in journals

Vasconcelos, C. C. de, Watanabe, E., & Leoncio, W. (2018). The impact of attorneys on judicial decisions: empirical evidence from civil cases. *International Journal for Court Administration*, **9(2)**.

Chapters in books

Leoncio, W., & Wiberg, M. (2018). Evaluating equating transformations from different frameworks. In *Springer Proceedings in Mathematics & Statistics*, eds. M. Wiberg, S. Culpepper, R. Janssen, J. González, D. Molenaar, pp. 101–109, **Vol. 233**.

Conference presentations

Leoncio, W., Wiberg, M., (2017). Assessing equating transformations in IRT and kernel equating methods. (poster) *2017 International Meeting of the Psychometric Society (IMPS)*, Zürich, Switzerland, 18/7/2017 to 21/7/2017.

References

Prof. Michela Battauz
Department of Economics and Statistics
University of Udine
Via Tomadini, 30/A-33100, Udine, Italy
Phone: +39 0432 249581
e-mail: michela.battauz@uniud.it

Prof. Marie Wiberg
Department of Statistics, Umeå University
Samhällsvetarhuset, Biblioteksgränd 6, Umeå universitet, 901 87 Umeå, Sweden
Phone: +46 90 786 95 24
e-mail: marie.wiberg@umu.se