Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
Scienze Statistiche
Ciclo XXIX

Settore Concorsuale di afferenza: 13/D1
Settore Scientifico disciplinare: SECS-S/01

ITEM RESPONSE THEORY EQUATING WITH THE NON-EQUIVALENT GROUPS WITH COVARIATES DESIGN

Presentata da: Valentina Sansivieri

Coordinatore dottorato	**Relatore**
Chiar.ma Prof.ssa Alessandra Luati	Chiar.ma Prof.ssa Stefania Mignani

Esame finale anno 2017

Preface

A test is an assessment intended to measure an examinee's knowledge (Kolen and Brennan, 2014). Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen and Brennan, 2014). Beginning in the early 1980s there was an increasing attention to equating mainly attributable to the increase in the number and variety of testing programs that use multiple forms of tests. In fact, scores on tests often are used as one piece of information in making important decisions: for example, an agency or institution might need to decide what test score is required to certify individuals for a profession or to admit students into a college, university, or the military. Making decisions in these contexts requires that test are administered on multiple occasions. For example, college admissions tests are administered on particular days, referred to as *test dates*, so examinees can have some flexibility in choosing when to be tested. Tests also are given over many years to track educational trends over time. If the same test questions are administered on each test date, then examinees might inform others about the test questions. Or, an examinee who tested twice might be administered the same test questions on the two test dates (Kolen and Brennan, 2014).

This work mainly deals with equating using Item Response Theory (IRT) (Lord, 1980). Equating using IRT (or IRT equating) typically is a three-step process. First, item parameters are estimated with a computer software. Second, parameter estimates are scaled to a base IRT scale using a linear transformation. Third, if number-correct scoring is used, number-correct scores on the new form are converted to the number-correct scale on an old form and then to scale scores (Kolen and Brennan, 2014). The main purpose of this work is to introduce covariates in IRT equating to reduce the standard error of equating (SEE). In general, equating with covariates is motivated by an interest in the possibility of using information about the examinees' background to increase the precision of the estimators. Bränberg and Wiberg (2011) and Wiberg and Bränberg (2015) have described methods for performing, respectively, observed score linear equating and kernel equating using covariates, while IRT equating with covariates has not been developed yet.

The idea to use covariates in IRT equating is to insert them in the calculation of the probability $P(y_{ji})$, where y_{ji} denote the response of examinee j to item i, using the IRT with Covariates (IRT-C) model (Tay et al., 2015), which is a single-class/restricted Mixed-Measurement IRT with Covariates (MM-IRT-C) approach (Tay et al., 2011). In particular, it is explained how to conduct IRT true-score equating (Lord, 1980) and IRT observed-score equating (Lord, 1980) using the IRT-C model.

IRT true-score equating is based on the concept of *test characteristic curve* of a test, which is defined as the sum of the probabilities $P(y_{ji})$ on the items of the test and which is considered to be the true score on the test. The method, substantially, consists in finding the equivalent true scores on the equated forms (Lord, 1980). Kolen and Brennan (2014) developed this method using the

three parameter logistic (3PL) IRT model; here this method is described using the IRT-C model.
IRT observed score equating uses the IRT model to produce an estimated distribution of observed number-correct scores on each form, which are then equated using equipercentile methods (Lord, 1980). Kolen and Brennan (2014) developed this method using the recursion formula (Lord and Wingersky, 1984) to obtain the observed score distribution for examinees of a given ability; here this method is described modifying the recursion formula so it is obtained the observed score distribution for examinees of a given ability and covariate.
Summarizing, the main contributions of this work are the development of the IRT true-score equating with covariates and the IRT observed-score equating with covariates. Both the methods will be discussed together with simulations and a real application.

Structure of the thesis The first chapter will provide a general overview of equating. The first section will define the basic concepts of equating while the second one will describe the practical issues.
The second chapter will illustrate traditional equating methods, using different designs. Then it will describe IRT equating methods and the kernel method of test equating. Then it will explain how to average equating functions and how to conduct equating with small samples. Subsequently it will focus on standard errors of equating. Finally, it will describe some recent ideas in test equating.
The third chapter will focus on equating with covariates. In particular, it will be explained how to use covariates in observed score linear equating, kernel equating and in a Bayesian context.
The fourth chapter will start describing the MM-IRT-C model and its precursor models. Subsequently it will describe how to conduct IRT equating using the MM-IRT-C model. Then a simulation study is used to examine the properties of the proposed equating methods. Finally, a real data example is given to show how the methods can be used in practice.

Acknowledgements

I would like to thank Prof. Stefania Mignani and Mariagiulia Matteucci for their supervision and for encouraging me to do my best.

I am deeply grateful to Prof. Marie Wiberg, for making possible my period of study in the Department of Statistics at the Umeå University (Sweden). I especially thank her for her patience and precious advices.

Finally, I would like to thank my family and everyone who loves me.

Contents

Chapter 1

Basic Concepts and Practical Issues

1.1 Basic Concepts

This section provides a general overview of equating. It briefly describes the concept of equating, the process of scaling, the list of steps for implementing equating, the properties of equating, the data collection designed used in equating and the equating error.

1.1.1 Definition of Equating

Test specifications provide guidelines for developing the test (some of the material in this chapter is based on Kolen and Brennan (2014) and von Davier (2011)). In some situations it can be useful to administer a different collection of test items, referred to as a *test form*, to examinees who are tested on different test dates. A test form is a set of test items that is built according to content and statistical test specifications: the test developers ensure that these test forms are as similar as possible to one another in content and statistical characteristics. A *link* between scores on two tests is a transformation from a score on one test to a score on another test. Equating is the strongest form of linking between the scores on two tests. Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably.

1.1.2 The Context of Equating

There are at least three reasons to have multiple forms of a test (and consequently equating). The first is security. Many testing programs administer high-stakes examinations in which performance has an important impact upon the examinee and the public: conferring a license or certificate to practice a profession, permitting admittance to a college or other training program, or granting credit for an educational experience. For a test score to have validity in any of these circumstances, it is crucial that it reflects the uncontaminated knowledge and ability of the examinees. Therefore, security is a concern and it is often desirable to give different forms to examinees seated beside each other, those who take the examination on different days, or those who take the examination on more than one occasion (Linn, 1989).

A second and related reason for different test forms is the current movement to open testing. Many programs find it necessary or desirable to release test items to the public (Holland and Rubin, 1982a). When this occurs, it is not possible to use the released items on future forms of a test without providing examinees an unfair advantage. For example, the Italian National Evaluation Institute for the School System (Invalsi) must release test items to the public because it is required by Italian law.

A third reason for different forms is that test content, and therefore test questions, by necessity change gradually over time. Knowledge in virtually all occupations and professions evolves and it is crucial for the test to reflect the current state of practice. For example, it is obvious that today's medical licensure and certification examinations should include questions on HIV and AIDS, whereas these topics were not relevant in the past. Even when the knowledge does not so obviously change, the context within which test items are presented is at risk of becoming dated. One could imagine a clinical scenario in medicine where descriptions of a patient's condition should be rewritten to include current drugs; in law one might want to include references to timely cases and rulings, especially if they lead to different interpretations of the law; in an educational context, Invalsi might want to include items about contemporary Italian authors in the Italian test.

Summing up, equating can be applied in all of the contexts when it is necessary to use multiple test forms. Nowadays, international surveys such as the Programme for International Student Assessment (PISA) use equating to compare multiple test forms over time (Gebhardt and Adams, 2007). PISA is a worldwide study by the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations of 15-year-old school pupils' scholastic performance on mathematics, science, and reading. It was first performed in 2000 and then repeated every three years. Finally, even if equating was born in an educational context, we can find publications about it in several contexts: see, for example, Kyoung Yim and Sun (2006) and Camilli et al. (1995).

1.1.3 Within-Grade (Horizontal) Equating and Across-Grade (Vertical) Equating in an educational context

Two primary situations exist for equating multiple test forms: the horizontal equating and the vertical equating situations (Lissitz and Huynh, 2003). Horizontal equating is designed to test different groups of students that are assumed to be at approximately the same level. It occurs within grade, where multiple forms are used to test the same general content. An example of the horizontal equating situation is the case in which a school system has a test for graduation and students are allowed to retake the test if they fail. The retakes are on different forms of the test that are all equated to provide comparable scores. The cut-off for failing is set at the same score level no matter how often a student retakes the test, thus ensuring a constancy of the standard for passing from administration to administration. The table of specifications for each test is also the same, thus ensuring that content is comparable and that the dimensions of knowledge that underlie the test are the same in each case. The difficulty level will be approximately the same for each form of the test, as well.

Vertical equating may be used when testing students who are at different levels of education. It entails across-grade testing of the same general content. A classic example of the vertical equating situation is that of a mathematics test that is used to track expertise across middle school.

In this scenario, the tests at different grade levels are of differing content, but still focus on the same general concept, say, mathematics ability. The students are expected to show performance improvements at each year, and these improvements should be reflected in a steady increase in their ability to do mathematics. The tests for two subsequent grades should be linked so that scores are directly comparable along a common continuum or dimension. The content must have some sense of commonality across grades in order to be meaningfully equated across grade levels. These scales are often considered developmental, in the sense that they encourage the examination of changes in a students score across grades that indicate the improvement in that students competency level. Sometimes the equating is only for adjacent grades and sometimes equating is across the whole school experience.

1.1.4 Equating and Score Scales

The numbers that are associated with examinee performance on educational or psychological tests are defined through the *process of scaling.* This process produces a *score scale*, and the scores that are reported to examinees are referred to as *scale scores.* A key component in the process of developing a score scale is the *raw score* for an examinee on a test, which we indicate with X.
The *summed score*, X, is defined as

$$X = \sum_{i=1}^{n} V_i^*, \tag{1.1}$$

where n is the number of items on the test and V_i^* is a random variable indicating score on item i. Equation 1.1 is often used as a raw score when all of the items on a test are of the same format. The weighted summed score, X_w,

$$X_w = \sum_{i=1}^{n} w_i V_i^*, \tag{1.2}$$

uses weights, $0 \leq w_i \leq 1$, to weight the item score for each item. Various procedures for choosing weights include choosing weights so that each item contributes a desired amount to the raw score. With alternate test forms, the raw scores typically do not have a consistent meaning across forms. For this reason, scores other than raw scores are used as primary score scales. The raw score is transformed to a scale score. For summed scores, the scale score, M_X, is a function of the summed score, X, such that $M_X = M_X(X)$. This transformation often is provided in tabular form. For weighted summed scores, the scale score, M_{Xw}, is a function of the weighted summed score, X_w, such that $M_{Xw} = M_{Xw}(X_w)$. Linear or nonlinear transformations of raw scores are used to produce scale scores that can be meaningfully interpreted. Normative, score precision, and content information can be incorporated (Kolen and Brennan, 2014, pp.394-420).

Incorporating Normative Information

Thorndike (1971) defined *norms* as descriptive statistics that are compiled to permit the comparison of a particular score with the scores earned by the members (or groups of members) of some defined population.

Incorporating normative information begins with the administration of the test to a norm group. After we calculate the mean and standard deviation of the raw scores for the norm group we can specify the mean and standard deviation of the scale scores to use this transformation:

$$sc(y) = \frac{\sigma(sc)}{\sigma(Y)} y + \left[\mu(sc) - \frac{\sigma(sc)}{\sigma(Y)}\mu(Y)\right], \tag{1.3}$$

where $\mu(Y)$ and $\sigma(Y)$ are the mean and standard deviations of raw scores in the norm group on form Y and $\mu(sc)$ and $\sigma(sc)$ are the desired mean and standard deviation of the scale scores.

Incorporating Score Precision Information

Scale score units should be of the most appropriate order of magnitude to express their accuracy of measurement. The use of too few score points fails to preserve all of the information contained in raw scores; however, the use of too many scale score points might lead test users to attach significance to scale score differences that are predominantly due to measurement error. By making suitable assumptions, the approximate range of scale scores that produces the desired score scale property is

$$6 \frac{h^*}{z_\gamma \sqrt{1-\rho_{XX'}}}, \tag{1.4}$$

where h^* is the width of the desired confidence interval, z_γ is the unit-normal score associated with the confidence coefficient γ and $\rho_{XX'}$ is test reliability (Kolen and Brennan, 2014, pp.402-403).

Incorporating Content Information

In item mapping we define *the response probability (RP) level* as the probability of a correct response that is associated with mastery for all items on a test, expressed as a percentage. The *mastery level* for a specific item is defined as the scale score for which the probability times 100 of correctly answering the item equals the *RP* level. Given the overall *RP* level, the mastery level for each item in a set of items can be found. Each item is mapped to a particular point on the score scale that represents the item's mastery level. The mastery level for items can be found by regressing probability of correct response on scale score, using procedures such as logistic regression, or by using an Item Response Theory (IRT) model. Items are usually reported in item maps only if they discriminate well between examinees who score above and below the score. The outcome of the item mapping procedure is the specification of test items that represent various scale score points.

Another way to incorporate content information is to use *scale anchoring.* The first step in scale anchoring is to develop an item map. Then, a set of scale score points is chosen, such as a selected set of percentiles. Subject-matter experts review the items that map near each of the selected points and develop general statements that represent the skills of the examinees scoring at each point. *Standard setting procedures* are used to find the scale score points that differentiate between adjacent levels (von Davier, 2011, pp.47-48).

A multistep process is used to put scores from a new test onto an existing score-reporting scale (which is built using one of the previous methods). Before the new test form is administered, there exists a conversion, $s(y)$, for an old test form that takes raw scores, y, on the old test form Y onto the score-reporting scale. This old-form scaling function, $s(y)$, is independent of the new form. Once data are collected on the new form, data from the new form and the old form are used to compute a raw-to-raw equating function, $eq(x)$, that links raw scores x on a new test X to those of an old test form Y. The final step in the process is to produce a function that converts the equated X raw scores to the score-reporting scale by composing the equating function, $y = eq(x)$ with $s(y)$. This puts the raw scores of X onto the reporting scale, $s(eq(x))$.

1.1.5 Equating in practice

According to (Kolen and Brennan, 2014, p.7), the steps for implementing equating are:

1. *Decide on the purpose for equating.*

2. *Construct alternate forms.*

3. *Choose a design for data collection.* Equating requires that data are collected for providing information on how the test forms differ statistically.

4. *Implement the data collection design.* The test is administered and the data are collected as specified by the design.

5. *Choose one or more operational definitions of equating.* Equating requires that a choice is made about what types of relationships between forms are to be estimated.

6. *Choose one or more statistical estimation methods.*

7. *Evaluate the results of equating.*

1.1.6 Properties of Equating

Defining desiderable equating properties is an essential step, because these properties have been used as the principal basis for developing equating procedures. Some properties focus on individuals' scores, others on distributions of scores. At the individual level, ideally, an examinee taking one form would earn the same reported score regardless of the form taken. At the distribution level, for a group of examinees, the same proportion would earn a reported score at or below, say, x on Form X as they would on Form Y.

Symmetry Property

The symmetry property (Lord, 1980) requires that the equating must be the same regardless of which test is labeled X and which is labeled Y.

Same Specifications Property

Test forms must be built to the same content and statistical specifications if they are to be equated (Kolen and Brennan, 2014, p.9).

Equity Properties

True score is the score that an examinee would has earned had there been no measurement error (Kolen and Brennan, 2014, p.8). Lord's equity property requires that for every true score T, the conditional frequency distribution of the equating function $eq(x)$ given T must be the same as the conditional frequency distribution of y (Lord, 1980, p.195). To make the description of this property more precise, define Form X as the new form, X as the random variable score on Form X, and x as a particular score on Form X; Form Y as the old form, Y as the random variable score on Form Y, and y as a particular score on Form Y; G as the cumulative distribution of scores on Form Y for the population of examinees; $eq(x)$ as an equating function that is used to convert scores on Form X to the scale of Form Y; and G^* as the cumulative distribution of $eq(x)$ for the same population of examinees. Formally:

$$G^*[eq(x) \mid T] = G(y \mid T), \quad \forall T, \tag{1.5}$$

Lord's equity property specified in equation 1.5 is possible only if Form X and Form Y are essentially identical. However, identical forms typically cannot be constructed in practice.
A less restrictive version of Lord's equity property is the *first-order equity property* or *weak equity property*, under which examinees with a given true score have the same mean converted score on Form X as they have on Form Y (Holland and Rubin, 1982b). Define E as the expectation operator, an equating achieves the first-order equity property if

$$E[eq(X) \mid T] = E(Y \mid T), \quad \forall T. \tag{1.6}$$

Finally, we have the *second-order equity property* which is said to hold to the extent that the conditional standard errors of measurement, after equating, are similar for the alternate forms (Kolen and Brennan, 2014, p.320).

Observed Score Equating Properties

For the *equipercentile equating property* (Thorndike, 1971), the converted scores on Form X have the same distribution as scores on Form Y. Formally:

$$G^*[eq(x)] = G(y). \tag{1.7}$$

Under the *mean equating property*, converted scores on the two forms have the same mean. Under the *linear equating property*, converted scores on the two forms have the same mean and standard deviation. When the equipercentile equating property holds, the linear and the mean equating must also hold. When the linear equating property holds, the mean equating property also must hold.

Invariance across groups

Under the invariance across groups property (Lord, 1980), eq_Y must be the same regardless of the population from which it is determined.

1.1.7 Data Collection Designs Used in Test Score Equating

There are a number of data collection designs which can be used.

Single Group Design

The single-group design is the simplest data collection design. In the singlegroup design, all examinees in a single sample of examinees from population P take both tests. The single-group design can provide accurate equating results with relatively small sample sizes, but in most equating situations, it is impossible to arrange for enough testing time for every examinee to take more than one test. The weaknesses of this design are the *order effects*: what if Form X was administered first to all examinees followed by Form Y? If fatigue was a factor in examinee performance, then Form Y could appear more difficult than Form X because examinees would be tired when administered Form Y; on the other hand, if familiarity with the test increased performance, then Form Y could appear to be easier than Form X.

Counterbalancing Design

The counterbalancing design is a single group design where one-half of the test booklets are printed with Form X following Form Y, and the other half are printed with Form Y following Form X. In packaging, test booklets having Form X first are alternated with test booklets having Form Y first. When the test booklets are handed out, the first examinee takes Form X first, the second examinee takes Form Y first, the third examinee takes Form X first, and so on. This spiraling process helps to ensure that the examinee group receiving Form Y first is comparable to the examinee group receiving Form X first. If the effect of taking Form X after taking Form Y is the same as the effect of taking Form Y after taking Form X, then the equating relationship will be the same between the forms taken first as it is between the forms taken second. Otherwise, a differental order effect is said to have occurred, and the equating relationship would differ. In this case, the data for the form that is taken second might need to be disregarded, which could lead to instability in the equating. Other practical problem is that because two forms must be administered to the same examinees, testing time needs to be doubled, which often is not practically feasible. The primary benefit in using this design is that it typically has small sample size requirements.

Equivalent (or Random) Groups Design

In this design, examinees are randomly assigned the form to be administered. A spiraling process is one procedure that can be used to randomly assign forms using this design. In one method for spiraling, Form X and Form Y are alternated when the test booklets are packaged. When the test booklets are handed out, the first examinee takes Form X, the second examinee takes Form Y, the third examinee takes Form X, and so on. This spiraling process typically leads to randomly equivalent groups taking Form X and Form Y. When using this design, the difference between group-level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms. One practical feature of this design is that each examinee takes only one form of the test, thus minimizing testing time relative to a design in which examinees take more than one form. In addition, more than one new form can be equated at the same time by including the additional new forms in the spiraling process. Finally, this design can also be used over time if we assume randomness. The weaknesses of this design are: it requires that all the forms be available and administered at the same time; large sample sizes are typically needed; practical issues should be considered (for example, if examinees were systematically seated boy-girl, boy-girl, then it is possible that boys get administered Form X and the girls Form Y).

Common-Item Nonequivalent Groups Design (or Nonequivalent Groups with Anchor Test (NEAT) Design)

In this design, Form X and Form Y have a set of items in common, and different groups of examinees are administered the two forms. The set of items in common is called "anchor test". This design has two variations. When the score on the set of common items contributes to the examinee's score on the test, the set of common items is referred to as *internal.* The internal common items are chosen to represent the content and statistical characteristics of the old form. For this reason, internal common items typically are interspersed among the other items in the test form. Because the items in an internal anchor test count towards the score, examinees are unlikely to skip them. On the other hand, once anchor test items have been used in the test administration of the old form, the items may become susceptible to security breaches and become known by examinees taking the new form to be equated. The primary problems with internal anchor tests are context effects, which can occur when common items are administered in different locations (e.g., common Item 10 in one form is Item 20 in the other form) or under different testing conditions (i.e., paper and pencil versus computer delivered), or when they are adjacent to different kinds of items in the two tests.

When the score on the set of common items does not contribute to the examinee's score on the test, the set of common items is referred to as *external*. For best practices, it is important to disguise the external anchor test so that it appears to be just another section of the test. One reason for this is that some examinees may identify the anchor test and, knowing that it does not count towards their final score, skip it or use the time to work on sections that count towards their score. It is best practice to exclude from the equating analysis any examinees whose anchor test performance is inconsistent with their total test performance.

The set of common items should be a *mini version* of the total test form (Kolen and Brennan, 2014, p.18). Each common item should occupy a similar location (item number) in the two forms and it should be exactly the same (no wording changes) in the old and new forms.

This design requires that only one test form be administered per test date; in contrast, the equivalent groups design typically requires different test form to be administered to equivalent subgroups of examinees, and the single group design requires that more than one form be administered to each examinee. The flexibility of this design has a cost: the larger the differences between examinee groups, the more difficult it becomes for the statistical methods to separate the group and form differences.

Nonequivalent Groups with Covariates (NEC) Design

In the NEC design (Wiberg and Bränberg, 2015) a sample of examinees from population P are administered test form X, and another sample of examinees from population Q are administered test form Y. For both samples we also have observations on background variables correlated with the test scores (i.e., covariates). In the last step, a synthetic population is defined and an equating is performed on this population (Andersson et al., 2013).

1.1.8 Error in Estimating Equating Relationships and Evaluation of the Results

Estimated equating relationships typically contain estimation error. A major goal in designing and conducting equating is to minimize such equating error. *Random equating error* is present whenever samples from populations of examinees are used to estimate parameters that are involved in estimating an equating relationship. Random error is typically indexed by the *standard error of equating*, which is the standard deviation of score equivalents over replications of the equating procedure. As the sample size becomes larger, the standard error of equating becomes smaller. Random error can be controlled by using large samples of examinees, by choosing an equating design that reduces such error, or both.

Systematic equating error results from violations of the assumptions and conditions of equating and it is difficult to quantify. Over time, after a large number of test forms are involved in the scaling and equating process, both random and systematic errors tend to accumulate.

After the equating is conducted, the results should be evaluated. Such evaluation requires that criteria for equating be identified. Estimating random error using standard error of equating can be used to develop criteria. Criteria for evaluating equating can also be based on consistency of results with previous results. The properties of equating can also be used to develop evaluative criteria.

1.2 Practical Issues

In this section the practical issues that are involved in conducting equating are described. The test development process, the data collection and the quality control procedures are discussed in all their facets.

1.2.1 Test Specifications

The *content specifications* are developed by considering the purpose of testing, and they provide an operational definition of the content that the test is intended to measure (Kolen and Brennan, 2014, pp.285–286). The content specifications typically include the content areas to be measured and the item types to be used, with the number of items per content area and item types to be used, with the numbers of items per content area and item types specified precisely. A test form must be sufficiently long to be able to achieve the purpose of the test, and it must provide a large enough sample of the domain for the alternate forms to be similar: in general, the lenght of a test depends on the purposes of testing, the heterogeneity of the content measured, and the nature of the test specifications.
Statistical specifications often are based on classical statistics such as the mean, standard deviation, and distribution of item difficulties and discriminations for a particular group of examinees.

1.2.2 Characteristics of Common-Item Sets

In constructing common-item sections, the sections should be long enough to represent test content adequately (Kolen and Brennan, 2014, pp.287–289). The number of common items to use should consider both content and statistical grounds. For example, because educational tests tend to be heterogeneous, larger numbers of common items are likely required for equating to be adequate in practice. Experience suggests the rule of thumb that a common item set should be at least 20% of the length of a total test containing 40 or more items, unless the test is very long, in which case 30 common items might suffice.

1.2.3 Changes in Test Specifications

Test specifications often evolve over time. For example, a particular content area might become obsolete and be replaced by a new area. With small changes, however, testing programs often continue to attempt to equate, often with only minimal problems. When the test specifications are modified significantly, scores obtained before the test was modified cannot be considered interchangeable with scores obtained after the test was modified, even if an equating process is attempted (Kolen and Brennan, 2014, pp.286–287).

1.2.4 Choosing Among Data Collection Equating Designs

As can be seen in Table 1.1, the choice of a data collection design requires making a series of trade-offs. The equivalent groups design typically results in the fewest test development complications,

Table 1.1: Comparison of Data Collection Designs.

Design	Test Administration Complications	Test Development Complications	Statistical Assumptions Required
Equivalent Groups	Moderate-more than one form needs to be spiraled	None	Minimal-that random assignment to forms is effective
Single Group with Counterbalancing	Major-each examinee must take two forms and order must be counterbalanced	None	Moderate-that order effects cancel out and random assigment is effective
Common-Item Nonequivalent Groups	Make sure anchor test is not obvious for examinees	Representative common-item sets need to be constructed	Stringent-that common items measure the same construct in both groups, the examinee groups are similar, and required statistical assumptions hold
Nonequivalent Groups with Covariates	None	Relevant Covariates need to be collected	Covariates capture the difference between the groups

because there is no need to develop common-item sets that are representative of the content of the total test. If not enough examinees are available for using the equivalent groups design, then the single group design might be preferable. When only a single form can be administered on a test date and equating is to be conducted, the choice of equating design is restricted to a design that uses common items. It can happen that you may not have access to common items, thus you need to use the NEC design.

1.2.5 Developing Equating Linkage Plans

When conducting equating, a choice is made about which old form or forms are to be used for equating a new form or forms (Kolen and Brennan, 2014, pp. 292–300). The choice of the old form or forms has a significant effect on how random and systematic equating error affects score comparisons across forms. Constructing equating plans can be very complicated. A particular form might need to be ruled out as an old form because of security concerns or because many of the examinees to be included in the equating were administered the old form on a previous occasion. Also, an old form might be found to have bad items (e.g., items that are ambiguous or negatively discriminating), which could rule out its use in equating.

One procedure that is often used to help solve the problems associated with developing linkage plans is to use two old forms to equate new forms. This process is referred to as *double linking.* In applying double linking, the new forms are equated separately to each of the old forms. The resulting equating relationships could then be averaged. The process of double linking has several strengths:

- it provides a built-in stability check on the equating process;
- it suggests problems with statistical assumptions, quality control, administration or security;
- it provides a great equating stability;
- it can reduce random error equating.

To construct equating linkage plans for the common-item non equivalent groups design with internal common items, the following four rules can be used:

1. Minimize the number of links that affect the comparison of scores on forms given at successive times;
2. Use links to the same time of the year as often as possible;
3. Minimize the number of links connecting each form back to the initial form;
4. Avoid linking back to the same form too often.

Obviously, all of these rules cannot be followed simultaneously when constructing a plan that uses single links. Choosing a plan involves a series of compromises that must be made in the context of the testing program under consideration.

Single Link Plan in Figure 1.1 might be a reasonable compromise: in this plan, rule 1 is followed reasonably closely; rule 2 is followed for nearly 1/2 of the forms; rules 3 and 4 are followed reasonably closely.

Double linking is useful in common-item nonequivalent groups design because it provides a built-in check on the equating process leading to greater equating stability; in contrast, when using the common-item nonequivalent groups design, double linking requires that two sets of common items which are content representative be used in the development of new forms, which sometimes can be difficult.

1.2.6 Examinee Groups Used in Equating

Equating relationships typically are group dependent: for this reason, more adequate equating is expected when the examinees used in the equating study are as similar as possible to the entire group that is tested (Harris, 1993). In the common-item nonequivalent groups design large differences between the old and the new groups can cause significant problems in estimating equating relationships, both for traditional and IRT equating methods. Large groups differences can lead to failure of the statistical assumptions for any equating method to hold.

A consideration when conducting equating is wheter or not to eliminate examinees who have taken the test previously. One argument for removing examinees who are repeating the test is that they might have seen the old form or common items, which could bias the equating. In contrast, excluding repeating examinees reduces sample size, which might lead to inadequate equating precision.

Other considerations are whether to delete examinees whose scores are very low or who omitted many items and wheter to eliminate test centers or testing sessions that had administration problems.

1.2.7 Sample Size Requirements

The smaller the sample size, the more restricted is the class of stable equating methods. The best practices solution to the small sample size problem may be to report raw scores and state that they cannot be compared across test forms. Another option may be to make strong assumptions about the equating function (Kolen and Brennan, 2014, p.303).

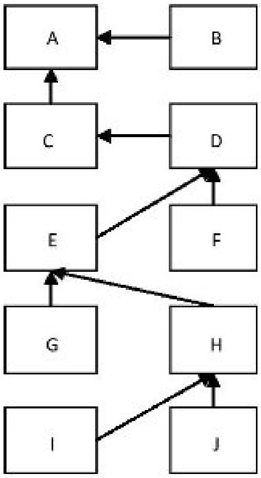


Figure 1.1: An hypothetical single link plan.
Source: (Kolen and Brennan, 2014, p.297).

1.2.8 Test Administration and Standardization Conditions

The conditions under which a test is administered should be standardized in order for tests administered at different locations and at different times to be comparable to one another (Kolen and Brennan, 2014, pp.331–333). Some issues related to standardization that could have significant effects on equating include the following:

1. *Changes in the number of items on the test.*

2. *Changes in timing of the test.*

3. *Changes in motivation conditions.*

4. *Security breaches.* Examinees are found to have had prior exposure to test forms or items that appear in the forms involved in the equating, which suggests that a security breach occurred.

5. *Scrambling of test items for security purposes.* Sometimes, test items within forms are scrambled to discourage examinee copying;

6. *Changes in the font used in printing the test or in the pagination used.*

7. *Use of calculators.* If calculators are allowed in some administrations and not in others, then scores from administrations that allow calculators are not directly comparable to scores from administrations that do not allow calculators.

1.2.9 Quality Control

Quality control checks are vital to adequate equating (Kolen and Brennan, 2014, p.333). Some of the quality control checks that can be made are the following:

1. *Check that the test administration conditions are followed properly.*

2. *The answer keys are correctly specified.*

3. *The equating procedures that are specified are followed correctly.*

4. *The score distributions and score statistics are consistent with those observed in the past.*

Chapter 2

Equating Models and Evaluation of the results

At the beginning of this chapter we illustrate traditional equating methods, using the equivalent groups and the NEAT designs. Then we describe item response theory equating methods and the kernel method of test equating. Then we expain how to average equating functions and how to conduct equating with small samples. Subsequently we focus on standard errors of equating. Finally, we describe some recent ideas in test equating.

2.1 Equivalent groups

The focus of this section is on the methods that are designed to achieve the observed score equating properties and these methods are developed in the context of the equivalent groups design (the single group design is not recommended, because order typically has a negative effect on scores).

2.1.1 Traditional Equating Methods

Traditional methods include mean equating, linear equating and equipercentile equating.

Mean Equating

Define Form X as the new form, let X represent the random variable score on form X, and let x represent a particular score on Form X; and define Form Y as the old form, let Y represent the random variable score on Form Y, and let y represent a particular score on Form Y. Also, define $\mu(X)$ as the mean on Form X and $\mu(Y)$ as the mean on Form Y for a population of examinees. In this method, scores on the two forms that are an equal distance away from their respective means are set equal:

$$x - \mu(X) = y - \mu(Y), \tag{2.1}$$

from which we obtain:

$$m_Y(x) = y = x - \mu(X) + \mu(Y), \tag{2.2}$$

in which $m_Y(x)$ refers to a score x on Form X transformed to the scale of Form Y using mean equating.

Linear Equating

Define $\sigma(X)$ and $\sigma(Y)$ as the standard deviations of Form X and Form Y scores, respectively. In linear equating, standardized deviation scores on the two forms are set equal:

$$\frac{x - \mu(X)}{\sigma(X)} = \frac{y - \mu(Y)}{\sigma(Y)}, \tag{2.3}$$

from which we obtain:

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + [\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)]. \tag{2.4}$$

It's simple to verify that:

$$E[m_Y(x)] = \mu(Y), \tag{2.5}$$
$$E[l_Y(x)] = \mu(Y), \tag{2.6}$$
$$\sigma[l_Y(x)] = \sigma(Y). \tag{2.7}$$

Therefore, the mean and standard deviation of the Form X scores equated to the Form Y scale are equal to the mean and standard deviation, respectively, of the Form Y scores.

Equipercentile Equating

The function e_Y is defined to be the equipercentile equating function in the population if

$$G^* = G, \tag{2.8}$$

where G is the cumulative distribution function of Y in the population and G^* is the cumulative distribution function of e_Y in the same population.
When X and Y are continuous random variables, the equipercentile equating function is: (Holland and Rubin, 1982a)

$$e_Y(x) = G^{-1}[F(x)], \tag{2.9}$$

where F is the cumulative distribution function of X in the population and G^{-1} is the inverse of the cumulative distribution function G.
To be an equating function, e_Y must be symmetric. By this symmetry property we obtain:

$$e_X(y) = F^{-1}[G(y)], \tag{2.10}$$

where F^{-1} is the inverse of the cumulative distribution function F. In equation 2.10 we find the equipercentile equating function for converting Form Y scores to the Form X scale.
To conduct equipercentile equating in practice, let n_X represent the number of items on Form X of a test. Define X as a random variable representing test scores on form X that can take on the integer values $0, 1, ..., n_X$. Define $f(x)$ as the discrete density function and $F(x)$ as the discrete cumulative distribution function. Consider a possible noninteger value of x. Define x^* as the integer that is closest to x such that $x^* - .5 \leq x < x^* + .5$. The percentile rank function for Form X is

$$\begin{aligned} P(x) &= 100\{F(x^*-1) + [x-(x^*-.5)][F(x^*) - F(x^*-1)]\}, \\ &\quad -.5 \leq x < n_X + .5, \\ &= 0, \quad x < .5, \\ &= 100, \quad x \geq n_X + .5. \end{aligned} \tag{2.11}$$

To find the inverse of the percentile rank function, which is referred to as the percentile function, solve equation 2.11 for x. For a given percentile rank P^*, the percentile is

$$\begin{aligned} x_U(P^*) = P^{-1}[P^*] &= \frac{P^*/100 - F(x_U^* - 1)}{F(x_U^*) - F(x_U^* - 1)} + (x_U^* - .5), \quad 0 \leq P^* < 100, \\ &= n_X + .5, \quad P^* = 100. \end{aligned} \tag{2.12}$$

In equation 2.12, for $0 \leq P^* < 100$, x_U^* is the smallest integer score with a cumulative percent $100[F(x)]$ that is greater than P^*. An alternative expression for the percentile is

$$\begin{aligned} x_L(P^*) = P^{-1}[P^*] &= \frac{P^*/100 - F(x_L^*)}{F(x_L^* + 1) - F(x_L^*)} + (x_L^* + .5), \quad 0 < P^* \leq 100, \\ &= -.5, \quad P^* = 0. \end{aligned} \tag{2.13}$$

In equation 2.13, for $0 < P^* \leq 100$, x_L^* is the largest integer score with a cumulative percent $100[F(x)]$ that is less than P^*. In equipercentile equating, the interest is in finding a score on Form Y that has the same percentile rank as a score on Form X. Referring to y as a score on Form Y, let n_Y refer to the number of items on Form Y, let $g(y)$ refer to the discrete density of y, let $G(y)$ refer to the discrete cumulative distribution of y, let $Q(y)$ refer to the percentile rank of y, and let Q^{-1} refer to the inverse of the percentile rank function for Form Y. Then the Form Y equipercentile equivalent of score x on Form X is

$$e_Y(x) = y = Q^{-1}[P(x)], \quad -.5 \leq x \leq n_X + .5. \tag{2.14}$$

To find $e_Y(x)$ given by equation 2.14 use

$$\begin{aligned} e_Y(x) &= \frac{P(x)/100 - G(y_U^* - 1)}{G(y_U^*) - G(y_U^* - 1)} + (y_U^* - .5), \quad 0 \leq P(x) < 100, \\ &= n_Y + .5, \quad P(x) = 100. \end{aligned} \tag{2.15}$$

In equation 2.15, for $0 \leq P(x) < 100$, y_U^* is the smallest integer score with a cumulative percent $100[G(y)]$ that is greater than $P(x)$.

Conducting equipercentile equating using equation 2.15 always results in equated scores in the range $-.5 \leq e_Y(x) < n_Y + .5$. In equipercentile equating the equated scores on Form X would have the same distribution as the scores on Form Y: however, when test scores are discrete, it isn't true.

Estimating Observed Score Equating Relationships

The methods have been described using population parameters, but in practice sample statistics are substituted for the parameters. One estimation problem is how to calculate the function P^{-1} when the frequency at some score points is zero. A procedure is to add a small relative frequency to each score, and then adjust the relative frequency so they sum to one:

$$\hat{g}_{adj}(y) = \tfrac{\hat{g}(y) + adj}{1 + (n_Y + 1)adj},$$

where $\hat{g}(y)$ is the relative frequency that was observed and adj is our small quantity (Kolen and Brennan, 2014, p.46).

2.1.2 Smoothing in Equipercentile Equating

When sample percentiles and percentile ranks are used to estimate equipercentile relationships, equating is not sufficiently precise because of sampling error. Smoothing methods have been developed that produce estimates of the equipercentile relationships more precise than the unsmoothed relationships.
Define x_i as a particular score on Form X and $e_Y(x_i)$ as the population equipercentile equivalent at that score, and define $\hat{e}_Y(x_i)$ as the sample estimate. Also assume that $E[\hat{e}_Y(x_i)] = e_Y(x_i)$. Define $\hat{t}_Y(x_i)$ as an alternative estimator of $e_Y(x_i)$ that results from using a smoothing method and assume that $E[\hat{t}_Y(x_i)] = t_Y(x_i)$.
Total error can be partitioned into random error $(\hat{t}_Y(x_i) - t_Y(x_i))$ and systematic error $(t_Y(x_i) - e_Y(x_i))$ components as follows:

$$\hat{t}_Y(x_i) - e_Y(x_i) = [\hat{t}_Y(x_i) - t_Y(x_i)] + [t_Y(x_i) - e_Y(x_i)]. \tag{2.16}$$

In terms of squared quantities we have:

$$\begin{aligned} mse[\hat{t}_Y(x_i)] &= E[\hat{t}_Y(x_i) - e_Y(x_i)]^2 = E[\hat{t}_Y(x_i) - t_Y(x_i)]^2 + [t_Y(x_i) - e_Y(x_i)]^2 \\ &= var[\hat{t}_Y(x_i)] + \{bias[t_Y(x_i)]\}^2. \end{aligned} \tag{2.17}$$

Smoothing methods are designed to produce smooth functions which contain less random error than that for unsmoothed equipercentile equating, but they can introduce systematic error: smoothing at score point x_i is useful if $mse[\hat{t}_Y(x_i)]$ is less than $var[\hat{e}_Y(x_i)]$.
A smoothing method should improve estimation and it should be flexible enough to handle many distributions and equipercentile relationships. Finally, there should be a statistical framework for studying fit.

Presmoothing Methods

Presmoothing methods are used to smooth the score distributions. One important property is moment preservation: it means that the smoothed distribution has at least some of the same central moments as the observed distribution.

Polynomial Log-Linear Method fits polynomial functions to the log of the sample density: (Darroch and Ratcliff, 1972)

$$log[N_X f(x)] = \omega_0 + \omega_1 x + \omega_2 x^2 + ... + \omega_{C^*} x^{C^*}. \tag{2.18}$$

The ω parameters in this model can be estimated by the method of maximum likelihood. The resulting fitted distribution has the moment preservation property that the first C^* moments of the fitted distribution are identical to those of the sample distribution. For assessing the fit of these models, likelihood ratio chi-square goodness-of-fit statistics are calculated for each C^* that is fit and are tested for significance. In addition, because the models are hierarchical, likelihood ratio difference chi-squares can be tested for significance. In model selection, the simplest model that adequately fits the distribution might be preferred.

Strong True Score Methods require the use of a parametric model for true scores. The *beta4 method* is one of these methods and it results in a smooth distribution of observed scores (Lord, 1965). In the development of the beta4 procedure, the true score distribution, $\psi(\tau)$, is assumed to be four-parameter beta and the conditional distribution of observed score given true score, $f(x \mid \tau)$, is assumed to be either binomial or compound binomial. The observed score distribution that results from the use of the following equation:

$$f(x) = \int_0^1 f(x \mid \tau)\psi(\tau)\, d\tau, \tag{2.19}$$

is referred as the *beta4 distribution*. For estimating this distribution and the associated true score distribution we can use the method of the moments. One important property of these method is that the first four central moments of the fitted distribution agree with those of the sample distribution. The fit of the model can be evaluated by using statistical methods, such as a standard chi-square goodness-of-fit. The degrees of freedom are the number of score points ($K+1$, to account for a score of 0), minus 1, minus the number of parameters fit.

Postsmoothing Methods

In postsmoothing methods, the equipercentile equivalents, $\hat{e}_Y(x)$, are smoothed directly. The method to be described makes use of cubic smoothing splines (Kolen, 1984). For each integer score, x_i, the spline function is

$$\hat{d}_Y(x) = v^*_{0i} + v^*_{1i}(x - x_i) + v^*_{2i}(x - x_i)^2 + v^*_{3i}(x - x_i)^3, x_i \leq x < x_i + 1. \tag{2.20}$$

The weights $v^*_{0i}, v^*_{1i}, v^*_{2i}, v^*_{3i}$ change from one score point to the next, so that there is a different cubic equation defined between each integer score. At each score point,x_i, the cubic spline is continuous. The spline is fit over the range of scores x_{low} to x_{high}, $0 \leq x_{low} \leq x \leq x_{high} \leq n_X$, where x_{low} is the lower integer score in the range and x_{high} is the upper integer score in the range. Defined O as a parameter which controls the degree of smoothing, the function, over score points, satisfies the following constraint:

$$\frac{\sum_{i=low}^{high} \left[\frac{\hat{d}_Y(x_i) - \hat{e}_Y(x_i)}{\hat{se}[\hat{e}_Y(x_i)]} \right]^2}{x_{high} - x_{low} + 1} \leq O. \tag{2.21}$$

In this equation, the term $\hat{se}[\hat{e}_Y(x_i)]$ is the estimeted standard error of equipercentile equating. The use of the standard error results in the smoothed and unsmoothed relationships being closer when the standard error is small, and allows them to be farther apart when the standard error is large. The parameter $O \geq 0$ is set by the investigator and its values between 0 and 1 produce adequate results in practice.

2.2 NEAT design

In this section the methods that are designed to achieve the observed score equating properties are developed in the context of the NEAT design.

2.2.1 Linear Methods with the NEAT design

Denote the common-item set and the random variable score on the common-item set as V. V can be an internal or an external set of common items. Assume that X and V are taken by a group of examinees from Population 1, and Y and V are taken by a group of examinees from Population 2. An equating function is viewed as being defined for a single population. For this reason, Populations 1 and 2 must be combined to obtain a single population for defining an equating relationship. In the *synthetic population* Populations 1 and 2 are weighted by w_1 and w_2, respectively, where $w_1+w_2=1$ and $w_1, w_2 \geq 0$.

For the NEAT design, the linear equation for equating observed scores on X to the scale of observed scores on Y is:

$$l_{Ys}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y), \tag{2.22}$$

where s indicates the synthetic population. The four synthetic population parameters in equation 2.22 can be expressed in terms of parameters for Populations 1 and 2 as follows:

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X), \tag{2.23}$$

$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y), \tag{2.24}$$

$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2, \tag{2.25}$$

$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2, \tag{2.26}$$

where the subscripts 1 and 2 refer to Populations 1 and 2, respectively. For the common-item nonequivalent groups design, X is not administered to examinees in Population 2, and Y is not administered to examinees in Population 1. Therefore, $\mu_2(X)$, $\sigma_2^2(X)$, $\mu_1(Y)$, and $\sigma_1^2(Y)$ in equations 2.23 - 2.26 cannot be estimated directly. The methods considered next make different statistical assumptions in order to express these four parameters as functions of directly estimable parameters.

Tucker Method

The Tucker method makes two types of assumptions in order to estimate the parameters in equations 2.23 - 2.26 that cannot be estimated directly (Gulliksen, 1950).

The first type of assumption concerns the regressions of total scores on common-item scores. First, the regression of X on V is assumed to be the same linear function for both Populations 1 and 2. A similar assumption is made for Y on V. For X and V the regression assumptions are:

$$\alpha_2(X \mid V) = \alpha_1(X \mid V), \tag{2.27}$$

and

$$\beta_2(X \mid V) = \beta_1(X \mid V), \tag{2.28}$$

where $\alpha_1(X \mid V) = \frac{\sigma_1(X,V)}{\sigma_1^2(V)}$ and $\beta_1(X \mid V) = \mu_1(X) - \alpha_1(X \mid V)\mu_1(V)$ are the slope and the intercept, respectively, for the regression of X on V in Population 1 and they are directly observed; $\alpha_2(X \mid V) = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}$ and $\beta_2(X \mid V) = \mu_2(X) - \alpha_2(X \mid V)\mu_2(V)$ are the slope and the intercept, respectively, for the regression of X on V in Population 2 and they are not directly observed.
For Y and V the regression assumptions are:

$$\alpha_1(Y \mid V) = \alpha_2(Y \mid V), \tag{2.29}$$

and

$$\beta_1(Y \mid V) = \beta_2(Y \mid V). \tag{2.30}$$

The second type of assumption concern the conditional variances of total scores given common-item scores: the conditional variance of X given V is assumed to be the same for Populations 1 and 2; a similar statement holds for Y given V. These assumptions are:

$$\sigma_2^2(X)[1 - \rho_2^2(X,V)] = \sigma_1^2(X)[1 - \rho_1^2(X,V)], \tag{2.31}$$

and

$$\sigma_1^2(Y)[1 - \rho_1^2(Y,V)] = \sigma_2^2(Y)[1 - \rho_2^2(Y,V)], \tag{2.32}$$

where ρ is a correlation and the quantities that are not directly observable are to the left of the equalities.

The synthetic population means and variances in equations 2.23 - 2.26 can be show to be: (Kolen and Brennan, 2014, pp.107-108)

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)], \tag{2.33}$$

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)], \tag{2.34}$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_1^2[\mu_1(V) - \mu_2(V)]^2, \tag{2.35}$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) - w_1\gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_2^2[\mu_1(V) - \mu_2(V)]^2, \tag{2.36}$$

where the $\gamma - terms$ are the regression slopes

$$\gamma_1 = \alpha_1(X \mid V) = \frac{\sigma_1(X,V)}{\sigma_1^2(V)}, \tag{2.37}$$

$$\gamma_2 = \alpha_2(X \mid V) = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}. \tag{2.38}$$

The Tucker linear equating function is obtained by using the results from equations 2.33-2.38 in equation 2.22.

Levine Observed Score Method

This method uses equation 2.22 to related observed scores on X to the scale of observed scores on Y (Levine, 1955). However, the assumptions for this method pertain to true scores T_X, T_Y and T_V:

$$X = T_X + E_X, \tag{2.39}$$

$$Y = T_Y + E_Y, \tag{2.40}$$

$$V = T_V + E_V, \tag{2.41}$$

where E_X, E_Y, and E_V are errors that have zero expectations and are uncorrelated with true scores. The Levine method assumes that X, Y, and V are all measuring the same thing:

$$\rho_1(T_X, T_V) = \rho_2(T_X, T_V) = 1, \tag{2.42}$$

and

$$\rho_1(T_Y, T_V) = \rho_2(T_Y, T_V) = 1. \tag{2.43}$$

The regression of T_X on T_V is assumed to be the same linear function for both Populations 1 and 2, and a similar assumption is made for the regression of T_Y on T_V. The slope of T_X on T_V is $\alpha_1(T_X \mid T_V) = \frac{\rho_1(T_X,T_V)\sigma_1(T_X)}{\sigma_1(T_V)}$, but, using the equation 2.42, we have $\alpha_1(T_X \mid T_V) = \frac{\sigma_1(T_X)}{\sigma_1(T_V)}$. Similarly, $\alpha_2(T_X \mid T_V) = \frac{\sigma_2(T_X)}{\sigma_2(T_V)}$. Using these results, the assumption of equal true score regression slopes for T_X on T_V in Populations 1 and 2 is:

$$\frac{\sigma_2(T_X)}{\sigma_2(T_V)} = \frac{\sigma_1(T_X)}{\sigma_1(T_V)}. \tag{2.44}$$

In the same manner, for the slopes of T_Y on T_V:

$$\frac{\sigma_1(T_Y)}{\sigma_1(T_V)} = \frac{\sigma_2(T_Y)}{\sigma_2(T_V)}. \tag{2.45}$$

The assumption of equal true score regression intercepts for T_X on T_V in Populations 1 and 2 is:

$$\mu_2(X) - \frac{\sigma_2(T_X)}{\sigma_2(T_V)}\mu_2(V) = \mu_1(X) - \frac{\sigma_1(T_X)}{\sigma_1(T_V)}\mu_1(V). \tag{2.46}$$

Similarly, for the intercepts of T_Y on T_V:

$$\mu_1(Y) - \frac{\sigma_1(T_Y)}{\sigma_1(T_V)}\mu_1(V) = \mu_2(Y) - \frac{\sigma_2(T_Y)}{\sigma_2(T_V)}\mu_2(V). \tag{2.47}$$

The Levine method also assumes that the measurement error variance for X is the same for Populations 1 and 2. A similar assumption is made for Y and V. The error variance assumptions are:

$$\sigma_2^2(X) - \sigma_2^2(T_X) = \sigma_1^2(X) - \sigma_1^2(T_X), \tag{2.48}$$

$$\sigma_1^2(Y) - \sigma_1^2(T_Y) = \sigma_2^2(Y) - \sigma_2^2(T_Y), \tag{2.49}$$

and

$$\sigma_1^2(V) - \sigma_1^2(T_V) = \sigma_2^2(V) - \sigma_2^2(T_V). \tag{2.50}$$

It can be shown that the synthetic population means and variances in equations 2.23-2.26 are given by equations 2.33-2.34 with (Kolen and Brennan, 2014, p.112)

$$\gamma_1 = \frac{\sigma_1(T_X)}{\sigma_1(T_V)}, \tag{2.51}$$

and

$$\gamma_2 = \frac{\sigma_2(T_Y)}{\sigma_2(T_V)}. \tag{2.52}$$

The γ-terms in equations 2.51-2.52 can be estimated using the reliability of X, Y, and V as follows:

$$\gamma_1 = \frac{\sigma_1(X)\sqrt{\rho_1(X,X')}}{\sigma_1(V)\sqrt{\rho_1(V,V')}}, \tag{2.53}$$

and

$$\gamma_2 = \frac{\sigma_2(Y)\sqrt{\rho_2(Y,Y')}}{\sigma_2(V)\sqrt{\rho_2(V,V')}}. \tag{2.54}$$

Levine True Score Method

The following equation is used to equate true scores on X to the scales of true scores on Y:

$$l_{Ys}(t_X) = \frac{\sigma_s(T_Y)}{\sigma_s(T_X)}[t_X - \mu_s(X)] + \mu_s(Y). \tag{2.55}$$

Equations 2.33 and 2.34 are still valid for this method, with the γ-terms given by equations 2.51 and 2.52. Using the same assumptions about true scores discussed previously, we obtain the variances of T_X and T_Y:

$$\sigma_s^2(T_X) = \gamma_1^2 \sigma_s^2(T_V), \tag{2.56}$$

and

$$\sigma_s^2(T_Y) = \gamma_2^2 \sigma_s^2(T_V), \tag{2.57}$$

where

$$\sigma_s^2(T_V) = w_1\sigma_1^2(T_V) + w_2\sigma_2^2(T_V) + w_1 w_2[\mu_1(V) - \mu_2(V)]^2. \tag{2.58}$$

As with the Levine observed score method, $\sigma_1(X)\sqrt{\rho_1(X,X')}$ can be used for $\sigma_1(T_X)$, and corresponding expressions can be used for the other true score standard deviations. Then, given estimates of the required reliabilities, $l_{Y\,s}(t_X)$ in equation 2.55 can be determined.
From the equations 2.56 and 2.57, the slope of the equating relationship $l_{Y\,s}(t_X)$ in equation 2.55 is:

$$\frac{\sigma_s(T_Y)}{\sigma_s(T_X)} = \frac{\gamma_2}{\gamma_1}. \tag{2.59}$$

From equations 2.55, 2.59, 2.33 and 2.34, the intercept is:

$$\mu_s(Y) - (\frac{\gamma_2}{\gamma_1})\mu_s(X) = \mu_2(Y) - (\frac{\gamma_2}{\gamma_1})\mu_1(X) + \gamma_2[\mu_1(V) - \mu_2(V)]. \tag{2.60}$$

From equations 2.59 and 2.60 the linear equating relationship for Levine's true score method can be expressed as

$$l_Y(t_X) = (\frac{\gamma_2}{\gamma_1})[t_X - \mu_1(X)] + \mu_2(Y) + \gamma_2[\mu_1(V) - \mu_2(V)], \tag{2.61}$$

which gives the same Form Y equivalents as equation 2.55. From equation 2.61 we can note that Levine's true score method does not require the conceptual framework of a synthetic population (s does not appear as a subscript of l) and is invariant with respect to the weights w_1 and w_2.

In equations 2.55 and 2.61 observed scores are used in place of true scores. Levine's true score method applied to observed scores has a property called *First-Order Equity*: for the population of persons with a particular score on Y, the expected value of the linearly transformed scores on X

equals the expected value of the scores on Y, and these statement holds for all true scores on Y.In formal terms:

$$E[l_Y(X) \mid \psi^*(T_X) = \tau] = E[Y \mid T_Y = \tau] \quad \forall\tau, \tag{2.62}$$

where ψ^* is a function that relates true scores on X to true scores on Y.

Braun-Holland Linear Method

This method uses the mean and standard deviation which arise from using the frequency estimation assumptions to conduct linear equating (Holland and Rubin, 1982a). Under the assumption 2.69 we have:

$$\mu_s(X) = \sum_x x f_s(x), \tag{2.63}$$

and

$$\sigma_s^2(X) = \sum_x [x - \mu_s(X)]^2 f_s(x), \tag{2.64}$$

where $f_s(x)$ is taken from equation 2.77. The expressions in equation 2.63 and 2.64 can be substituted into equation 2.22: the resulting equation is the Braun-Holland linear method.

2.2.2 Equipercentile Methods with NEAT design

Define $f(x, v)$ as the joint distribution of total score and common-item score, $f(x)$ as the marginal distribution of scores on Form X, $h(v)$ as the marginal distribution of scores on the common items and $f(x \mid v)$ as the conditional distribution of scores on Form X for examinees earning a particular score on the common items. It can be shown that:

$$f(x \mid v) = \frac{f(x, v)}{h(v)}. \tag{2.65}$$

From equation 2.65, it follows that

$$f(x, v) = f(x \mid v)h(v). \tag{2.66}$$

Frequency Estimation Method

The distribution for the synthetic population are a weighted combination of the distributions for each population: (Thorndike, 1971)

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x), \tag{2.67}$$

and

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \tag{2.68}$$

where the subscript s refers to the synthetic population, the subscript 1 refers to the population administered Form X, and subscript 2 refers to the population administered Form Y. f and g refer to the distributions for Form X and Form Y, respectively, and w_1 and w_2 $(w_1 + w_2 = 1)$ are used to weight Populations 1 and 2 to form the synthetic population.
Direct estimates of $f_2(x)$ and $g_1(y)$ are unavailable: for this reason, statistical assumptions need to be invoked to obtain expressions for these functions. The assumptions made are:

$$f_1(x \mid v) = f_2(x \mid v) \tag{2.69}$$

and

$$g_1(y \mid v) = g_2(y \mid v), \tag{2.70}$$

which are true $\forall v$. From equation 2.66, the following equalities hold:

$$f_2(x, v) = f_2(x \mid v) h_2(v), \tag{2.71}$$

and

$$g_1(y, v) = g_1(y \mid v) h_1(v). \tag{2.72}$$

Combining the equalities in equations 2.71 and 2.72 with the assumptions in equation 2.69 and 2.70 we obtain:

$$f_2(x, v) = f_1(x \mid v)h_2(v) \tag{2.73}$$

and

$$g_1(y, v) = g_2(y \mid v)h_1(v). \tag{2.74}$$

The associated marginal distributions can be found as follows:

$$f_2(x) = \sum_v f_2(x, v) = \sum_v f_1(x \mid v)h_2(v), \tag{2.75}$$

and

$$g_1(y) = \sum_v g_1(y, v) = \sum_v g_2(y \mid v)h_1(v). \tag{2.76}$$

The expressions in equation 2.75 and 2.76 can be substituted into equation 2.67 and 2.68 to provide expressions for the synthetic populations as follows:

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x \mid v)h_2(v), \tag{2.77}$$

$$g_s(y) = w_1 \sum_v g_2(y \mid v)h_1(v) + w_2 g_2(y). \tag{2.78}$$

Define $F_s(x)$ and $G_s(y)$ as the cumulative distributions, respectively, of $f_s(x)$ and $g_s(y)$; P_s and Q_s as the percentile rank function, respectively, for Form X and Form Y; and P_s^{-1} and Q_s^{-1} as the percentile function, respectively, for Form X and Form Y.
The equipercentile function for the synthetic population is

$$e_{Ys}(x) = Q_s^{-1}[P_s(x)]. \tag{2.79}$$

Chained Equipercentile Equating

The Chained Equipercentile Equating involves the following steps: (Thorndike, 1971)

1. Find the equipercentile equating relationship ($e_{V1}(x)$) for converting scores on Form X to the common items based on examinees from Population 1.

2. Find the equipercentile equating relationship ($e_{Y2}(v)$) for converting scores on the common items to scores on Form Y based on examinees from Population 2.

3. To find the Form Y equipercentile equivalent of a Form X score, take

$$e_{Y(chain)} = e_{Y2}[e_{V1}(x)]. \tag{2.80}$$

The chained equipercentile equating has at least two theoretical shortcomings. First, this method involves equating a long test to a short test. Second, this method does not directly incorporate a synthetic population, so it is not clear for what population the relationship holds. However, the method might be useful in situations where the two groups differ, because it does not explicitly require that the two populations be very similar.

2.3 Equating with IRT

Equating using IRT is a three-step process: in the first step, item parameters are estimated; in the second step, parameter estimates are scaled to a base IRT scale using a linear transformation; in the third step, if number-correct scoring is used, number-correct scores on the new form are converted to the number-correct scale on an old form and then to scale scores.

2.3.1 Transformations of IRT Scales

In the *equivalent groups design*, the IRT parameters for Form X can be estimated separately from the parameters for Form Y: if the same scaling convention for ability is used in the separate estimations, then the parameter estimates for the two forms are assumed to be on the same scale without further transformation.
In the *counterbalancing design*, the parameters for all examinees on both forms can be estimate together: for this reason, the parameter estimates are assumed to be on the same scale.
When conducting equating with the *NEAT design*, the parameter estimates that result from IRT

parameter estimation procedures are often on different IRT scales (this is true because the examinees who took Form X are not considered to be equivalent to the examinees who took Form Y). When the IRT model holds, a linear equation can be used to convert IRT parameter estimates to the same scale. Define Scale I and scale J as three-parameter logistic IRT scales that differ by a linear transformation. Then the θ-values for the two scales are related as follows:

$$\theta_{Jj} = A^*\theta_{Ij} + B^*, \tag{2.81}$$

where A^* and B^* are constants in the linear equation and θ_{Jj} and θ_{Ij} are values of θ for individual j on Scale J and Scale I. The item parameters on the two scales are related as follows:

$$a_{Ji} = \frac{a_{Ii}}{A^*}, \tag{2.82}$$

$$b_{Ji} = A^*b_{Ii} + B^*, \tag{2.83}$$

$$c_{Ji} = c_{Ii}, \tag{2.84}$$

where a_{Ji}, b_{Ji}, and c_{Ji} are the item parameters for the item i on Scale J and a_{Ii}, b_{Ii}, and c_{Ii} are the item parameters for the item i on Scale I. We have:

$$\begin{aligned} A^* &= \frac{\sigma(b_J)}{\sigma(b_I)}, && (2.85a)\\ &= \frac{\mu(a_I)}{\mu(a_J)}, && (2.85b)\\ &= \frac{\sigma(\theta(b_J))}{\sigma(\theta(b_I))}, && (2.85c) \end{aligned}$$

$$\begin{aligned} B^* &= \mu(b_J) - A^*\mu(b_I), && (2.86a)\\ &= \mu(b_J) - A^*\mu(b_I). && (2.86b) \end{aligned}$$

The means $\mu(b_J)$, $\mu(b_I)$, $\mu(a_I)$, and $\mu(a_J)$ in these equations are defined over one or more items with parameters that are expressed on both Scale I and Scale J. The standard deviations $\sigma(b_J)$ and $\sigma(b_I)$ are defined over two or more items with parameters that are expressed on both Scale I and Scale J. The standard deviations $\sigma(\theta(b_J))$ and $\sigma(\theta(b_I))$ are defined over two or more examinees with parameters that are expressed on both Scale I and Scale J.

The *mean/sigma method* uses the means and standard deviations of the b-parameter estimates from the common items in place of the parameters in equations 2.85a and 2.86a (Marco, 1977).
The *mean/mean method* uses the mean of the a-parameter estimates for the common items in place of the parameters in equations 2.85b to estimate the A^*-constant and the mean of the b-parameter estimates for the common items in place of the parameters in equations 2.86a to estimate the B^*-constant (Loyd and Hoover, 1980). The values of A^* and B^* then can be substituted into equations 2.81-2.84 to obtained the rescaled parameter estimates (which are often referred to as being *calibrated*).
One potential problem with the methods considered so far arises when various combination of $a-$, $b-$ and $c-$parameter estimates produce almost identical item characteristic curves over the range of ability at which most examinees score. This problem arises because the scale conversion methods described so far do not consider all of the item parameter estimates simultaneously. As a response to this problem there are the so called *characteristic curve methods* (Haebara (1980) and Stocking and Lord (1983)). Note that, for ability Scales I and J,

$$p_{ji}(\theta_{Jj}; a_{Ji}, b_{Ji}, c_{Ji}) = p_{ji}\left(A^*\theta_{Ij} + B^*; \frac{a_{Ii}}{A^*}, A^*b_{Ii} + B^*, c_{Ii}\right), \tag{2.87}$$

for examinee j and item i. Equation 2.87 states that the probability that the examinees of a given ability will answer a particular item correctly is the same regardless of the scale that is used to report the scores. If estimates are used in place of the parameters in equation 2.87, then there is no guarantee that the equality will hold over all items and examinees for any A^* and B^*. This lack of equality is exploited by the characteristic curve methods.

Haebara Approach

The function used by Haebara (1980) to express the difference between the item characteristic curves is the sum of the squared difference between the item characteristic curves for each item for examinees of a particular ability. For a given θ_j we have:

$$Hdiff(\theta_j) = \sum_{i:V}\left[p_{ji}(\theta_{Jj}; \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - p_{ji}\left(\theta_{Jj}; \frac{\hat{a}_{Ii}}{A^*}, A^*\hat{b}_{Ii} + B^*, \hat{c}_{Ii}\right)\right]^2. \tag{2.88}$$

The estimation process proceeds by finding A^* and B^* that minimize the following criterion:

$$Hcrit = \sum_j Hdiff(\theta_j). \tag{2.89}$$

Stocking and Lord Approach

Stocking and Lord (1983) used the sum, over items, of the squared difference,

$$SLdiff(\theta_j) = \left[\sum_{i:V} p_{ji}(\theta_{Jj}; \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - \sum_{i:V} p_{ji}\left(\theta_{Jj}; \frac{\hat{a}_{Ii}}{A^*}, A^* \hat{b}_{Ii} + B^*, \hat{c}_{Ii} \right) \right]^2. \tag{2.90}$$

The estimation process proceeds by finding A^* and B^* that minimize the following criterion:

$$SLcrit = \sum_j SLdiff(\theta_j). \tag{2.91}$$

The approach to solving for A^* and B^* in equations 2.89 and 2.91 is a computationally intensive iterative approach.

2.3.2 Comparisons Among Scale Transformation Methods

Research comparing the characteristic curve methods to the mean/sigma and mean/mean methods has generally found that the characteristic curve methods produce more stable results than the mean/sigma and mean/mean methods (Hanson and Béguin (2002)). In addition, Ogasawara (2000) found that the mean/mean method was more stable than the mean/sigma method.

2.3.3 Equating

Using estimated IRT abilities results in several practical problems: the whole 0/1 response string, rather than the number-correct score, is used to estimate θ, and this could involve that examinees with the same number-correct score receive different estimated abilities; estimates of θ are difficult to compute and costly to obtain; the estimated θ-values with the three-parameter logistic model are subject to greater amounts of measurement error for high and low ability examinees than for middle ability examinees. For these reasons, tests are often scored number-correct.

2.3.4 IRT true score equating

Four requirements are important for equating two unidimensional test x and y that measure the same ability (Lord, 1980): equity, invariance across groups, symmetry and same specifications (see subsection 1.1.6). An equating of true scores can satisfy the listed requirements. Lord (1980) described true-score equating as follows. If test X and test Y are both measures of the same ability θ, then their number-right true scores are related to θ by their test characteristic functions:

$$\tau_X(\theta) = \sum_{i:X} p_i(\theta), \quad \tau_Y(\theta) = \sum_{i:Y} p_i(\theta), \tag{2.92}$$

where $p_i(\theta)$ are the item response curves. Equations 2.92 are parametric equations for the relation between $\tau_X(\theta)$ and $\tau_Y(\theta)$. A single equation for the relationship is found (in principle) by eliminating θ from the two parametric equations. In practice, this relationship can be estimated by using estimated item parameters to approximate the $p_i(\theta)$ and then substituting a series of arbitrary values of θ into 2.92 and computing $\tau_X(\theta)$ and $\tau_Y(\theta)$ for each θ. The resulting paired values define $\tau_X(\theta)$ as a function of $\tau_Y(\theta)$ (or vice versa) and constitute an equating of these true scores. Since $\tau_X(\theta)$ and $\tau_Y(\theta)$ are each monotonic increasing functions of θ, it follows that $\tau_X(\theta)$ is a monotonic increasing function of $\tau_Y(\theta)$.

Kolen and Brennan (2014) described true-score equating using the three parameter logistic IRT model. The number-correct true score on Form X that is equivalent to θ_j is defined as

$$\tau_X(\theta_j) = \sum_{i:X} p_{ji}(\theta_j; a_i, b_i, c_i), \tag{2.93}$$

The number-correct true score on Form Y that is equivalent to θ_j is defined as

$$\tau_Y(\theta_j) = \sum_{i:Y} p_{ji}(\theta_j; a_i, b_i, c_i), \tag{2.94}$$

Equations 2.93 and 2.94 are referred to as *test characteristic curves* for Form X and Form Y. True scores on Form X and Y are associated with a value of θ only over the following ranges:

$$\sum_{i:X} c_i < \tau_X < n_X, \quad \sum_{i:Y} c_i < \tau_Y < n_Y, \tag{2.95}$$

because, when using the three-parameter logistic IRT model, as θ approaches $-\infty$, p_{ji} approaches c_i and not 0.
For a given θ_j, true scores $\tau_X(\theta_j)$ and $\tau_Y(\theta_j)$ are considered to be equivalent. The Form Y true score equivalent of a given true score on Form X is

$$irt_Y(\tau_X) = \tau_Y(\tau_X^{-1}), \quad \sum_{i:X} c_i < \tau_X < n_X, \tag{2.96}$$

where τ_X^{-1} is defined as the θ_j corresponding to true score τ_X. Equation 2.96 implies that true score equating is a three-step process:

1. Specify a true score τ_X on Form X.
2. Find the θ_j that corresponds to that true score (τ_X^{-1}).
3. Find the true score on Form Y, τ_Y, that corresponds to that θ_j.

Whereas Step 1 and Step 3 are straightforward, Step 2 requires the use of the Newton-Raphson method. To apply the Newton-Raphson method, an initial value is chosen for θ, which is referred to as θ^-. Given $func(\theta)$ and $func'(\theta)$, which are, respectively, a function of the parameter θ and its first derivative, a new value for θ is calculated as

$$\theta^+ = \theta^- - \frac{func(\theta)}{func'(\theta)}. \tag{2.97}$$

Typically, θ^+ will be closer to the root of the equation than θ^-. The new value then is redefined as θ^-, and the process is repetead until θ^+ and θ^- are equal at a specified level of precision.
To apply this method to IRT true-score equating, let τ_X be the true score whose equivalent is to be found. From equation 2.93 it follows that θ_j is to be found such that the expression

$$func(\theta_j) = \tau_X - \sum_{i:X} p_{ji}(\theta_j; a_i, b_i, c_i), \tag{2.98}$$

equals 0. The Newton-Raphson method can be employed to find this θ_j using the first derivative of $func(\theta_j)$ with respect to θ_j, which is

$$func'(\theta_j) = -\sum_{i:X} p'_{ji}(\theta_j; a_i, b_i, c_i), \tag{2.99}$$

where $p'_{ji}(\theta_j; a_i, b_i, c_i)$ is the first derivative of $p_{ji}(\theta_j; a_i, b_i, c_i)$ with respect to θ_j and it is defined as:

$$p'_{ji}(\theta_j; a_i, b_i, c_i) = \frac{1.7a_i(1-p_{ji})(p_{ji}-c_i)}{(1-c_i)}, \tag{2.100}$$

where $p_{ji} = p_{ji}(\theta_j; a_i, b_i, c_i)$. The resulting expressions for $func(\theta_j)$ and $func'(\theta_j)$ are substituted into equation 2.98.

In practice, the true score equating relationship is used to convert number-correct observed scores on Form X to number-correct observed scores on Form Y, because the true scores of examinees are never known. Remembering that the range of possible true score on Form X is $\sum_{i:X} c_i < \tau_X < n_X$, a procedure is needed for converting Form X scores outside this range when using true score equating with observed scores. Lord (1980) and Kolen (1981) presented ad hoc procedures to handle this problem.
We first describe the Lord (1980) ad hoc procedure. Consider applying the IRT observed-score equating (see below) just to a hypothetical group of examinees all at ability level $\theta = -\infty$. The observed scores for such a group of examinees have a mean of $\sum_i c_i$ and a variance of $\sum_i c_i(1-c_i)$ (Lord, 1980, p.45). For x scores below $\sum_{i:X} c_i$, let us take the equating function $y(x)$ to be a linear function of x chosen so that both y and $y(x)$ have the same mean and also the same variance in our hypothetical subgroup of examinees. This means we shall use a conventional "mean and sigma" linear equation based on this subgroup of examinees. This equating requires that

$$\frac{y(x) - \sum_{i:Y} c_i}{\sqrt{\sum_{i:Y} c_i(1-c_i)}} = \frac{x - \sum_{i:X} c_i}{\sqrt{\sum_{i:X} c_i(1-c_i)}}, \tag{2.101}$$

From equation 2.101 we obtain

$$y(x) = \frac{\sqrt{\sum_{i:Y} c_i(1-c_i)}}{\sqrt{\sum_{i:X} c_i(1-c_i)}}(x - \sum_{i:X} c_i) + \sum_{i:Y} c_i. \tag{2.102}$$

We use 2.102 for test x scores below $\sum_{i:X} c_i$; we use true-score equating 2.92 above $\sum_{i:X} c_i$. From equation 2.102 we note that when $x = \sum_{i:X} c_i$, we find that $y(x) = \sum_{i:Y} c_i$. The 2.102 is a good practical solution to an awkward problem.
We now describe the Kolen (1981) ad hoc procedure. It is as follows:

1. Set a score of 0 on Form X equal to a score of 0 on Form Y.
2. Set a score of the sum of the c_i-parameters on Form X equal to the sum of the c_i-parameters on Form Y.
3. Use linear interpolation to find equivalents between these points.
4. Set a score of n_X on Form X equal to a score of n_Y on Form Y.

To formalize this procedure, define τ_X^* as a score outside the range of possible true scores, but within the range of possible observed scores. Equivalent are defined by the following equation:

$$\begin{aligned} irt_Y(\tau_X^*) &= \frac{\sum_{i:Y} c_i}{\sum_{i:X} c_i}\tau_X^*, \quad 0 \leq \tau_X^* \leq \sum_{i:X} c_i, \\ &= n_Y, \quad \tau_X^* = n_X. \end{aligned} \tag{2.103}$$

2.3.5 IRT observed score equating

IRT observed score equating uses the IRT model to produce an estimated distribution of observed number-correct scores on each form, which then are equated using equipercentile methods. This method requires explicit specification of the distribution of ability in the population of examinees.

Lord (1980) described the IRT observed score equating as follows. We start with an estimate $\hat{\psi}(\theta)$ of the distribution $\psi(\theta)$ of θ in some specific group. The actual distribution of $\hat{\theta}_o$ in the group is an approximation to $\psi(\theta)$. The distribution of observed score x for the specific group can be estimated by

$$\hat{\phi}_x(x) = \frac{1}{N}\sum_{o=1}^{O} \hat{\phi}_x(x \mid \hat{\theta}_o), \tag{2.104}$$

where $o = 1, ..., O$ indexes the examinees in the specified group. If $\hat{\psi}(\theta)$ is continuous the estimated $\phi(x)$ is obtained by

$$\hat{\phi}_x(x) = \int_{-\infty}^{+\infty} \hat{\phi}_x(x \mid \theta)\hat{\psi}(\theta)\, d\theta. \tag{2.105}$$

The function $\hat{\phi}_x(x \mid \hat{\theta}_o)$ is given in (Lord, 1980, p.45). Similar equation apply for test y. Furthemore, since x and y are independently distributed when θ is fixed, the joint distribution of scores x and y for the specified group is estimated by

$$\hat{\phi}(x, y) = \frac{1}{O}\sum_{o=1}^{O} \hat{\phi}_x(x \mid \hat{\theta}_o)\hat{\phi}_y(y \mid \hat{\theta}_o), \tag{2.106}$$

or by

$$\hat{\phi}(x, y) = \int_{-\infty}^{+\infty} \hat{\phi}_x(x \mid \theta)\hat{\phi}_y(y \mid \theta)\hat{\psi}(\theta)\, d\theta. \tag{2.107}$$

The integrand of 2.107 is the trivariate distribution of θ, x and y. Since θ determines the true scores $\tau_X(\theta)$ and $\tau_Y(\theta)$, this distribution also represents the joint distributions of the four variables $\tau_X(\theta)$, $\tau_Y(\theta)$, x and y. The joint distribution contains all possible information about the relation of x to y. A plausibile procedure is to determine the equipercentile relationship between x and y from 2.106 or 2.107 and to treat this as an approximate equating.

Kolen and Brennan (2014) described the IRT observed score equating as follows. Define $f_r(x \mid \theta_j)$ as the distribution of number-correct scores over the first r items for examinees of ability θ_j. Define $f_1(x = 0 \mid \theta_j) = (1 - p_{j1})$ as the probability of earning a score of 0 on the first item and $f_1(x = 1 \mid \theta_j) = p_{j1}$ as the probability of earning a score of 1 on the first item. For $r > 1$, the recursion formula is: (Lord and Wingersky, 1984)

$$\begin{aligned} f_r(x \mid \theta_j) &= f_{r-1}(x \mid \theta_j)(1 - p_{jr}), \quad x = 0, \\ &= f_{r-1}(x \mid \theta_j)(1 - p_{jr}) + f_{r-1}(x - 1 \mid \theta_j)p_{jr}, \quad 0 < x < r, \\ &= f_{r-1}(x - 1 \mid \theta_j)p_{jr}, \quad x = r. \end{aligned} \tag{2.108}$$

The recursion formula gives the observed score distribution for examinees of a given ability. To find the observed score distribution for examinees at each ability is found and then these are accumulated. When the ability distribution is continuous, then

$$f(x) = \int_{\theta} f(x \mid \theta)\psi(\theta)\, d\theta, \tag{2.109}$$

where $\psi(\theta)$ is the distribution of θ. To implement this procedure in practice, some method is needed to perform the integration in equation 2.109. We can approximate the integral in equation 2.109 as follows:

$$f(x) = \sum_{j=1}^{N} f(x \mid \theta_j)\psi(\theta_j). \tag{2.110}$$

When the ability distribution is discrete, then

$$f(x) = \frac{1}{N}\sum_{j=1}^{N} f(x \mid \theta_j). \tag{2.111}$$

To conduct observed score equating, observed score distributions are found for Form X and for Form Y. For example, assume that the characterization of the ability distribution associated with equation 2.110 is used. The following distributions could be specified using this equation:

1. $f_1(x) = \sum_j f(x \mid \theta_j)\psi_1(\theta_j)$ is the Form X distribution for Population 1.

2. $f_2(x) = \sum_j f(x \mid \theta_j)\psi_2(\theta_j)$ is the Form X distribution for Population 2.

3. $g_1(y) = \sum_j g(y \mid \theta_j)\psi_1(\theta_j)$ is the Form Y distribution for Population 1.

4. $g_2(y) = \sum_j g(y \mid \theta_j)\psi_2(\theta_j)$ is the Form Y distribution for Population 2.

These quantities then are weighted using synthetic weights described in Subsections 2.2.1 and 2.2.2 to obtain the distributions of X and Y in the synthetic population. Conventional equipercentile methods then are used to find score equivalents. Alternatives to the Lord-Wingersky algorithm can be found in González et al. (2016).

2.3.6 IRT True Score Versus IRT Observed Score Equating

Compared to IRT observed score equating, IRT true score equating has the advantages of (a) easier computation and (b) a conversion that does not depend on the distribution of ability (Kolen and Brennan, 2014, p.201). However, IRT true score equating has the disadvantage that it equates true scores, which are not available in practice. No justification exists for applying the true score relationship to observed scores.

IRT observed score equating has the advantage that it defines the equating relationship for observed scores. Also, assuming reasonable model fit, the distribution of Form X scores converted to the Form Y scale is approximately equal to the distribution of Form Y scores for the synthetic population of examinees. There is no theoretical reason to expect this property to hold for IRT true score equating.

Larger differences between IRT true and IRT observed score equating might be expected to occur near a number-correct score of all correct and near number-correct scores below the sum of the c-parameter estimates, because these are the regions where IRT true score equating does not produce equivalents. In practice, both methods should be applied with special attention paid to equating results near these regions.

Practical Caveat

Kolen and Brennan (2014) recommend the following when using IRT to conduct equating in practice:

1. Use both the Haebara and Stocking and Lord methods for scale transformation as well as the mean/sigma and mean/mean methods.

2. When equating number-correct scores, use both IRT true score equating and IRT observed score equating.

3. Whenever possible, conduct traditional equipercentile or linear methods on the forms that are being equated as a check.

Often all of the methods applied provide similar equating results and conversion to scale scores, which is reassuring. However, when the results for the different methods diverge, then a choice must be made about which results to believe. The assumptions required and the effects of poor parameter estimates need to be considered in these cases.

2.4 The Kernel Method of Test Equating

The kernel method of equating is a general procedure to equate one test form to another. This method of equating can be described in five steps (von Davier et al., 2004):

1. *Pre-smoothing.* In this step, estimates of the univariate and/or bivariate score probabilities are obtained by fitting statistical models to the raw data obtained by the data collection design.

2. *Estimation of the score probabilities.* In this step, a crucial role is played by the Design Function: it is a linear or nonlinear transformation of the estimated score distributions from Step 1 into the estimated score probabilities, $\hat{r}$ and $\hat{s}$, for test X and Y on the target population, T.

3. *Continuization.* In this step, we determine continuous approximations, $\hat{F}_{h_X}(x)$ and $\hat{G}_{h_Y}(y)$, to the estimated discrete cdf's, $\hat{F}(x)$ and $\hat{G}(y)$. Here we need to choose the *bandwidth parameters*, h_X and h_Y.

4. *Equating.* In this phase, the estimated equating function is formed from the two continuized cdf's, $\hat{F}_{h_X}(x)$ and $\hat{G}_{h_Y}(y)$, using:

$$\hat{e}_{h_X h_Y}(x) = \hat{G}_{h_Y}(\hat{F}_{h_X}(x)). \tag{2.112}$$

5. Calculating the standard error of equating.

In the following subsections we review the first four steps, and we wait to introduce the standard error of equating to a later point.

2.4.1 Pre-smoothing Using Log-Linear Models

To estimate a univariate score distribution using log-linear models, we will assume that the data come from an equivalent groups design (Holland and Thayer, 2000).
We denote N_j as the number of examinees in the sample with $X = x_j$; M_k as the number of examinees in the sample with $Y = y_k$; $N^* = \sum_j N_j$ and $M^* = \sum_k M_k$ as the two sample sizes. We can make the following assumption:
The vectors $\boldsymbol{N} = (N_1, ..., N_j)^t$ and $\boldsymbol{M} = (M_1, ..., M_k)^t$ are independent and they each have multinomial distributions:

$$\begin{aligned} Prob(\boldsymbol{N}) &= \frac{N^*!}{N_1!...N_j!} \prod r_j^{N_j}, \\ Prob(\boldsymbol{M}) &= \frac{M^*!}{M_1!...M_k!} \prod s_k^{M_k}, \end{aligned} \tag{2.113}$$

where $r_j = Prob\{X = x_j \mid T\}$ and $s_k = Prob\{Y = y_k \mid T\}$.
The log-likelihood function for $\boldsymbol{r}$ is:

$$L_r = \sum_j N_j log(r_j), \tag{2.114}$$

and for $\boldsymbol{s}$ is:

$$L_s = \sum_k M_k log(s_k). \tag{2.115}$$

To estimate the population parameters, we make this assumption:

The vector $\boldsymbol{r}$ satisfies a log-linear model

$$log(r_j) = \alpha + u_j + \boldsymbol{b_j^{*t}\beta}, \tag{2.116}$$

where $\boldsymbol{\beta}$ is a T_r-vector of free parameters, u_j is a known constant, α is the normalizing constant selected to make the sum of r_j equal to one and $\boldsymbol{b_j^*}$ is a T_r-vector of known constant. Maximum likelihood estimation proceeds maximizing the log-likelihood function given in 2.114. When a log-linear model of the form 2.116 is substituted for $logr_j$ in 2.114, L_r becomes a function of $\boldsymbol{\beta}$, $L_r(\boldsymbol{\beta})$, and can be maximized by solving $\frac{dL_r}{d\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. The solution, $\hat{\boldsymbol{\beta}}$, is the maximum likelihood estimate of $\boldsymbol{\beta}$. The maximum likelihood estimate of r_j is $\hat{r}_j = r_j(\hat{\boldsymbol{\beta}})$ (von Davier et al., 2004, pp.109-202).

2.4.2 Estimation of the Score Probabilities

The Design Function (DF) maps the population score probabilities for X and Y on the target population T. We describe the DF only for two designs (because we mainly talk about these designs in this text): equivalent groups and NEAT design (von Davier et al., 2004, pp.53-54).

Equivalent Groups Design. In this design, the DF is given by

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix} = DF(\boldsymbol{r}, \boldsymbol{s}) = \begin{pmatrix} \boldsymbol{I_J} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_K} \end{pmatrix} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix}, \tag{2.117}$$

where $\boldsymbol{I_J}$ is a JxJ identity matrix, $\boldsymbol{I_K}$ is a KxK identity matrix, $\boldsymbol{r}$ is the vector given by $(r_1, .., r_J)^t$ and $\boldsymbol{s}$ is the vector given by $(s_1, ..., s_K)^t$.

NEAT Design-Chain Equating. There are two levels of Design Functions that arise when the equating is carried out through Chain Equating in a NEAT Design. At the first level, there are two Design Functions that are from the two SG Designs inside the NEAT Design, denoted DF_P and DF_Q. DF_P is defined by

$$\begin{pmatrix} \boldsymbol{r_P} \\ \boldsymbol{t_P} \end{pmatrix} = DF_P(\boldsymbol{P}) = \begin{pmatrix} \boldsymbol{M_P} \\ \boldsymbol{N_P} \end{pmatrix} v(\boldsymbol{P}), \tag{2.118}$$

where $\boldsymbol{P}$ is the JxL matrix whose (j,l)-entry is $p_{jl} = Prob\{X = x_j, V = v_l \mid P\}, j = 1, ..., J, l = 1, ..., L$; $v(\boldsymbol{P}) = \begin{pmatrix} \boldsymbol{p_1} \\ \vdots \\ \boldsymbol{p_L} \end{pmatrix}$ where $\boldsymbol{p_l}$ denotes the lth column of $\boldsymbol{P}$; $\boldsymbol{M_P} = \overbrace{\begin{pmatrix} \boldsymbol{I_J}, & ..., & \boldsymbol{I_J} \end{pmatrix}}^{Ltimes}$; $\boldsymbol{N_P} = \overbrace{\begin{pmatrix} \boldsymbol{1_J^t} & \boldsymbol{0_J^t} & ... & \boldsymbol{0_J^t} \\ & & \vdots & \\ \boldsymbol{0_J^t} & ... & \boldsymbol{0_J^t} & \boldsymbol{1_J^t} \end{pmatrix}}^{Ltimes}$, where $\boldsymbol{1_J}$ is a (column) J-vector of $1's$ and $\boldsymbol{0_J}$ is a (column) J-vector of $0's$; $\boldsymbol{r_P} = \boldsymbol{M_P} v(\boldsymbol{P})$ and $\boldsymbol{t_P} = \boldsymbol{N_P} v(\boldsymbol{P})$.

DF_Q is defined by:

$$\begin{pmatrix} \boldsymbol{t_Q} \\ \boldsymbol{s_Q} \end{pmatrix} = DF_Q(\boldsymbol{Q}) = \begin{pmatrix} \boldsymbol{N_Q} \\ \boldsymbol{M_Q} \end{pmatrix} v(\boldsymbol{Q}), \tag{2.119}$$

where $\boldsymbol{Q}$ is the KxL matrix whose (k,l)-entry is $q_{kl} = Prob\{Y = y_k, V = v_l \mid Q\}, k = 1, ..., K, l = 1, ..., L$; $v(\boldsymbol{Q}) = \begin{pmatrix} \boldsymbol{q_1} \\ \vdots \\ \boldsymbol{q_L} \end{pmatrix}$ where $\boldsymbol{q_l}$ denotes the lth column of $\boldsymbol{Q}$ and N_Q, a $(LxKL)$ matrix, and M_Q, a $(KxKL)$ matrix, are defined analogously to M_P and N_P and $\boldsymbol{s_Q} = \boldsymbol{M_Q} v(\boldsymbol{Q})$ and $\boldsymbol{t_Q} = \boldsymbol{N_Q} v(\boldsymbol{Q})$. At the second level, these two Design Functions are combined into a single function

$$\begin{pmatrix} \boldsymbol{r_P} \\ \boldsymbol{t_P} \\ \boldsymbol{t_Q} \\ \boldsymbol{s_Q} \end{pmatrix} = DF(\boldsymbol{P}, \boldsymbol{Q}) = \begin{pmatrix} DF_P(\boldsymbol{P}) \\ DF_Q(\boldsymbol{Q}) \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \boldsymbol{M_P} \\ \boldsymbol{N_P} \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} \boldsymbol{N_Q} \\ \boldsymbol{M_Q} \end{pmatrix} \end{pmatrix} \begin{pmatrix} v(\boldsymbol{P}) \\ v(\boldsymbol{Q}) \end{pmatrix}. \tag{2.120}$$

NEAT Design- Poststratification Method. The Design Function for this design is given by

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix} = DF(\boldsymbol{P}, \boldsymbol{Q}, w) = \begin{pmatrix} r(\boldsymbol{P}, \boldsymbol{Q}, w) \\ s(\boldsymbol{P}, \boldsymbol{Q}, w) \end{pmatrix}, \tag{2.121}$$

where $0 \leq w \leq 1$ is a weight, $r(\boldsymbol{P}, \boldsymbol{Q}, w) = \sum_l \left[w + \frac{(1-w)t_{Ql}}{t_{Pl}} \right] \boldsymbol{p_l}$, $s(\boldsymbol{P}, \boldsymbol{Q}, w) = \sum_l \left[(1-w) + \frac{w(t_{Pl})}{t_{Ql}} \right] \boldsymbol{q_l}$ and t_{Pl} and t_{Ql} are column sums of P and Q, respectively.

2.4.3 Continuization

The cdf's of the score distributions for X and Y are defined as:

$$F(x) = \sum_{j, x_j \leq x} r_j, \tag{2.122}$$

$$G(y) = \sum_{k, y_k \leq y} s_k, \tag{2.123}$$

where $x, y \in \mathbb{R}$, These discrete cdf's have jumps at each score value, x_j or y_k. The Kernel smoothing continuizes a discrete random variable X by adding to it a continuous and independent random variable V^{**} with a positive constant h_X controlling the degree of smoothness. Let $X(h_X)$ denote the continuous approximation of X. Then

$$X(h_X) = X + h_X V^{**}. \tag{2.124}$$

The h_X is the bandwidth and is free to select to achieve certain practical purpose. Kernel function refers to the density function of V^{**}. Usually V^{**} is assumed to be a standard normal variable. Let $\Phi(z)$ denote the cdf of the Standard Normal and let h_X be a positive number. μ_X and σ_X^2 denote, respectively, the mean and the variance of X over T. Define a_X by

$$a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2}, \tag{2.125}$$

then the Gaussian kernel smoothing of the distribution of X has a cdf given by

$$F_{h_X}(x) = \sum_j r_j \Phi(R_{jX}(x)), \tag{2.126}$$

where

$$R_{jX}(x) = \frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}. \tag{2.127}$$

In a similar manner, G may be continuized using G_{h_Y} defined as

$$G_{h_Y}(y) = \sum_k s_k \Phi(R_{kY}(y)), \tag{2.128}$$

where

$$R_{kY}(y) = \frac{y - a_Y y_k - (1 - a_Y)\mu_Y}{a_Y h_Y}. \tag{2.129}$$

There are also some alternative kernels (von Davier, 2011, pp.159-174).
With reference to the selection of the "bandwidth" h_X (h_Y), there is a variety of ways to do it: the easiest is to always use a spefic fixed value (von Davier et al., 2004, pp.61-64).
Andersson and von Davier (2014) proposed to use

$$h_{adj} = .9\sigma_X N_X^{-1/5}, \tag{2.130}$$

as the choice for the bandwidth using a standard Gaussian kernel, where σ_X is the standard deviation of X and N_X is the sample size (Scott, 1992).
Liang and von Davier (2014) proposed a cross-validation method which uses one subsample to compute a set of Gaussian kernel densities (in this way there will be a density value at each score point for each given h) and the other subsample to compute the frequency k at each point: the density value from the first subsample and the frequency from the second subsample are plugged in a Poisson probability function as λ and k, respectively (in this way, for each h at each score point, there is a Poisson probability value) then the likelihood function is formed by multiplying the probability values at all score points and, finally, the h that corresponds to the largest likelihood is the optimal one.
Häggström and Wiberg (2014) proposed the use of a Double Smoothing (DS) procedure. First, start with a very smooth first estimate of the density function ($\hat{f}_{g_X}(x)$); second, improve on this first estimate by estimating f_{h_X} using $\hat{f}_{g_X}$ (in this way, we obtain $\hat{f}^*_{h_X}(x)$); finally, select the bandwidth value of h_X that minimizes the sum of the squared difference between the lth DS estimate $\hat{f}^*_{h_X}(x^*_l)$ and $\hat{r}^*_l$:

$$DS(h_X) = \sum_{l=1}^{2J-1} (\hat{r}^*_l - \hat{f}^*_{h_X}(x^*_l))^2, \tag{2.131}$$

where $\hat{r}^*_l = \hat{r}_{\frac{l+1}{2}}$ if l is odd, and $\hat{r}^*_l = \hat{f}^*_{h_X}(x^*_l)$ if l is even.

2.4.4 Equating

The KE function for equating X to Y on T is given by

$$\hat{e}_Y(x) = \hat{G}^{-1}_{h_Y}(\hat{F}_{h_X}(x)). \tag{2.132}$$

Analogously, the KE function for equating Y to X on T is given by

$$\hat{e}_X(y) = \hat{F}_{h_X}^{-1}(\hat{G}_{h_Y}(y)). \tag{2.133}$$

2.5 Further Details

In this section two practical problems are discussed: how to average results from more than one equating and how to conduct equating when the samples are too small to reliably conduct any classical equating procedures.

2.5.1 How to Average Equating Functions

There are many situations in which we can have multiple estimates of the same equating function for which averaging the different estimates may be appropriate. For example, in the NEAT design, several possible linear and nonlinear equating methods are available, thus it might be useful to average the results of some of the options for a final compromise method. We denote by $eq_1(x)$ and $eq_2(x)$ two different equating functions for linking scores on test X to scores on test Y. We will assume that $eq_1(x)$ and $eq_2(x)$ are strictly increasing continuous functions of x over the entire real line. Suppose it is desired to average $eq_1(x)$ and $eq_2(x)$ in some way, putting weight w on $eq_1(x)$ and $1 - w$ on $eq_2(x)$. Let $\oplus$ denote an operator that forms a weighted average of two such functions, eq_1 and eq_2, and puts weight w on eq_1 and $1 - w$ on eq_2. Our notation for any such weighted average of eq_1 and eq_2 is

$$weq_1 \oplus (1 - w)eq_2, \tag{2.134}$$

to denote the resulting equating function. We denote its value at some X-score, x, by

$$weq_1(x) \oplus (1 - w)eq_2(x). \tag{2.135}$$

If there are three such functions, eq_1, eq_2, eq_3, then their weighted average function is denoted as

$$w_1eq_1 \oplus w_2eq_2 \oplus w_3eq_3, \tag{2.136}$$

where the weights, w_i, sum to 1.

The operator $\oplus$ should possess various properties (von Davier, 2011, pp.90–92). There are different types of weighted average:

1. *The Point-Wise Weighted Average* of eq_1 and eq_2 is:

$$m(x) = weq_1(x) + (1 - w)eq_2(x), \tag{2.137}$$

   where w is a fixed value (von Davier, 2011, pp.92–93);

2. *The Angle Bisector Method* (Thorndike, 1971) says that if $eq_1(x)$ and $eq_2(x)$ are two linear equating functions that intersect at a point, then the linear function that bisects the angle between them is the point-wise weighted average

$$eq_{AB} = Weq_1 + (1 - W)eq_2, \tag{2.138}$$

   where

$$W = \frac{(1 + b_1^2)^{-1/2}}{(1 + b_1^2)^{-1/2} + (1 + b_2^2)^{-1/2}}, \tag{2.139}$$

   where b_1 and b_2 are the slopes of the linear equating functions $eq_1(x)$ and $eq_2(x)$, respectively (von Davier, 2011, pp.94–98);

3. *The symmetric w-average or swave* of two linear or nonlinear equating functions $eq_1(x)$ and $eq_2(x)$ (Holland and Strawderman, 1989) is:

$$eq_w(x) = weq_1(x - (1 - w)t(x)) + (1 - w)eq_2(x + wt(x)), \tag{2.140}$$

   where $t(x)$ is given by the equation

$$t(x) = eq_1(x - (1 - w)t) - eq_2(x + wt), \tag{2.141}$$

   which has a unique solution for t for any choice of x and w and for any strictly increasing continuous equating functions eq_1 and eq_2 (von Davier, 2011, pp.98–104).

2.5.2 Equating With Small Samples

Equating test scores is a statistical procedure, and its results, like those of most other statistical procedures, are subject to sampling variability. The smaller the samples of examinees from which the equating is computed, the more the equating results are likely to deviate from what they would be in a different pair of samples-or in the population that the samples represent.

One way to improve the accuracy of equipercentile equating in small samples of examinees is to presmooth the score distributions. However, if the samples of examinees are quite small, this technique may not reduce the sampling variability in the equating to an acceptable level. Another procedure that has been recommended is to establish a minimum sample size for equating.

A third approach is to estimate a relationship in a population on the basis of small-sample data is to use a strong model. Strong models require only a small number of parameters to be estimated from the data, in effect substituting assumptions for data. In test score equating, the strong model most commonly used is the linear equating model. However, when test forms differ in difficulty, the equating relationship between them typically is not linear and, if the difficulty difference is substantial, the relationship is not even approximately linear.

The Circle-Arc Method

Circle-arc equating is a strong model that does not assume the equating relationship to be linear (Divgi, 1987). The Circle-Arc Method constrains the equating curve to pass through two prespecified end points and an empirically determined middle point. In circle-arc equating, the lower end point corresponds to the lowest meaningful score on each form. On a multiple-choice test scored by counting the number of correct answers, the lowest meaningful score would typically be the expected score for an examinee who responds at random (e.g., without reading the items). The upper end point corresponds to the maximum possible score on each form. The middle point is determined by equating at a single point in the middle of the score distribution. If the middle point happens to lie on the line connecting the end points, that line is the estimated equating curve. If not, the next step is to use the end points and the middle point to determine the equating curve. There are two versions of circle-arc equating, and they differ in the way they fit a curve to these three points. We call one version *symmetric circle-arc equating* (Livingston and Kim, 2010) and the other *simplified circle-arc equating* (Livingston and Kim, 2009). Both methods are applications of the geometrical fact that if three points do not lie on a straight line, they uniquely determine a circle. Symmetric circle-arc equating fits a circle arc directly to the three data points. Simplified circle-arc equating transforms the three data points by decomposing the equating function into a linear component and a curvilinear component. The linear component is the line connecting the two end points. The curvilinear component is estimated by fitting a circle arc to the three transformed data points.
Figure 2.1 illustrates the simplified circle-arc procedure. The horizontal axis represents the score on the new form, that is, the test form to be equated. The vertical axis represents the corresponding score on the reference form. The two prespecified end points and the empirically determined middle point are indicated by the three small circles. The line connecting the two end points is the linear component of the estimated equating curve. The three data points are transformed by subtracting the y-value of that line, which we call $L(x)$. The three transformed points are indicated by the squares at the bottom of Figure 2.1. The middle point is transformed to a point above or below the horizontal axisabove it if the new form is harder than the reference form, and below it if the new form is easier than the reference form. In the example illustrated by Figure 2.1, the new form is harder than the reference form. The arc connecting the three transformed data points is shown at the bottom of Figure 2.1. This arc serves as an estimate of the curvilinear component of the equating function. For each possible raw score on the new form, there is a corresponding point on the arc. The next step is to add the linear component back in, by adding the height of the line $L(x)$ to the height of the arc. The three original data points are retransformed back to their original positions, and the full arc is transformed into an estimate of the equipercentile equating function, shown in the upper portion of the figure. The last step is to extend the equating transformation below the lower end point, by connecting that point linearly to the point corresponding to the minimum possible score on each form.

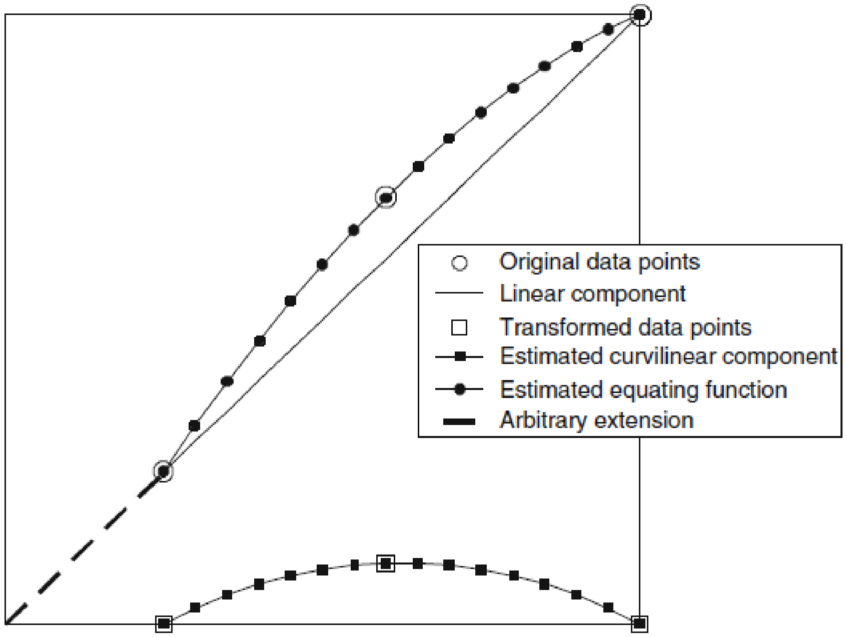

○ Original data points
— Linear component
□ Transformed data points
–■– Estimated curvilinear component
–●– Estimated equating function
▬ Arbitrary extension

Figure 2.1: Illustration of the simplified circle-arc equating method.
Source: (von Davier, 2011, p.113).

Nominal Weights Mean Equating

Babcock et al. (2012) proposed a new method called *Nominal Weights Mean Equating* as an alternative for dealing with very small samples. This method is a simplified version of Tucker linear equating (see subsection 2.2.1). We use the term nominal weights to distinguish from effective weights, which are typically based on the ratios of variances and/or covariances (Kolen and Brennan, 2014, pp.114-116). Nominal weights mean equating replaces covariance and variance terms with a simple ratio of number of items on the total test to the anchor test, making the weighting of the adjustment nominal instead of effective. We indicate the new test with X (Examinee Group 2 takes this form), the old test with Y (Examinee Group 1 takes this form), the anchor test with V (both groups take this form), the number of items in a given item set with n, the sample size of a given item set with N, the mean total test scores for each group with $\mu_2(X)$ and $\mu_1(Y)$ and the mean score on the common items for each group with $\mu_2(V_X)$ and $\mu_1(V_Y)$. The final score adjustment for nominal weights mean equating is

$$l_Y(x) = x - \mu_2(X) + \mu_1(Y) + \left[\frac{N_X n_Y + N_Y n_X}{[N_Y + N_X]n_V}\right][\mu_2(V_X) - \mu_1(V_Y)]. \tag{2.142}$$

There are two things to notice about Equation 2.142. First, the score adjustment is essentially a scaling up of the difference between the two groups performance on the equator items V. This makes a good set of equator items extremely important. Second, there are no variance or covariance terms in the equation. Depending on the context, this feature could be an advantage or a disadvantage compared with other methods. If the variances and covariances are well estimated, approximating the terms ignores valuable information that could be useful in making a score adjustment. If the variances and covariances are not well estimated, however, excluding them from the adjustment makes the adjustment less susceptible to equating error. This is exactly the case that methodologists deal with when in small-sample situations, making nominal weights mean equating a practical alternative.

A General Linear Method for Equating With Small Samples

Albano (2015) introduced a new approach to observed-score equating with small samples. Observed-score equating functions make different assumptions regarding the difficulty difference between forms and how this difference does or does not change across the X and Y scales. Based on these assumptions, form difficulty is then estimated using empirical score distributions for individuals taking X and Y. For simplicity, most of the discussion below is based on a situation where individuals taking X and Y are sampled from the same population. As a result, scores are presumed to come from an equivalent-groups or single-group equating design, where estimated differences in the X and Y score distributions reflect a combination of (a) actual differences in form difficulty, (b) random error caused by the sampling process, and (c) systematic error or bias caused by violations of the assumptions of a given function. The assumptions of a given function can be expressed in the form of a line within the coordinate space defined by X and Y. The ordered possible scores on X and their corresponding equated scores on Y make up the coordinates $(x, link_Y(x))$ for this equating line, where x is an observed score on X and $link_Y(x)$ a function linking or equating x to the Y scale.

A General Linear Function

Albano (2015) proposed a more flexible form of the identity function to take scale differences into account and after he observed that this identity line can also be shifted upward or downward so as to pass through any coordinate pair (β_X, β_Y). This results in the general linear equating function:

$$lin_Y(x) = y = \frac{\alpha_Y}{\alpha_X}x + \beta_Y - \frac{\alpha_Y}{\alpha_X}\beta_X, \tag{2.143}$$

where α_Y is an estimate of the variability in Y and α_X is an estimate of the variability in X.

Empirical Bayes Estimation

One way to improve an estimation process, especially when the data come from a small sample, is to incorporate collateral information (Efron and Morris, 1977). Collateral information for equating test scores is often available from equatings of other forms of the test we want to equate and of other tests. The idea of using collateral information suggests an empirical Bayes approach (Livingston and Lewis, 2009). The basic premise is that the current, small-sample equating can be regarded as a single observation randomly sampled from a large domain of possible equatings, each with its own new form, reference form, and samples of examinees. For a given pair of test forms, there is a *true* equating function: the function that would result from averaging over all the equatings of that particular pair of test forms with different samples of examinees. For a given raw score x on the new form, the empirical Bayes estimate of the corresponding equated score y on the reference form is a weighted average of a *current* estimate and a *prior* estimate, which we will call $y_{current}$ and y_{prior}. The current estimate is the equated score implied by the small-sample equating. The prior estimate is the average of the equating results in all the equatings used as collateral information, with the addition of the current equating. Formally:

$$\ddot{y}_{EB} = \frac{[var(y_{prior})]y_{current} + [var(y_{current})]y_{prior}}{var(y_{prior}) + var(y_{current})}. \tag{2.144}$$

This approach has three main problems:

1. Test forms differ in length: one solution to this problem could be to convert the scores on all the forms to percentages;
2. The Equation 2.144 requires an estimate of the sampling variance of the current equating, but it could be inaccurate because it is computed from a small-sample data;
3. It's difficult to decide what equatings to include as collateral information.

There is another simpler approach to using collateral information. For a given raw score x, if $y_{obs}(x)$ represents the equated score observed in the small-sample equating, the adjustment is simply

$$y_{adj}(x) = w[y_{obs}(x)] + (1 - w)x, \tag{2.145}$$

where w is a weight between 0 and 1. The practical question for implementing this procedure is how to choose a value for w: ideally, the value of w should vary with the size of the samples; the larger the samples, the greater the weight for the observed equating (Kim et al., 2008).

Introducing the New Form by Stages

Another, very different approach to the small-sample equating problem is to change the way in which new forms of the test are introduced (Puhan et al., 2009). This technique requires that the test be structured in testlets, small-scale tests that each represent the full test in content and format. It also requires that the test form given at each administration include one testlet that is not included in computing the examinees scores. With each new form, one of the scored testlets in the previous form is replaced. It is replaced in the printed new form by a new testlet, which is not scored. It is replaced in the scoring of the new form by the testlet that was new (and therefore was not scored) in the previous form. Each new test form is equated to the previous form in the group of examinees who took the previous form.

2.6 Evaluation of the results

The focus of the present section is on estimating random error, which is always present when the scores of examinees who are considered to be samples from a population or populations of examinees are used to estimate equating relationships. See Wiberg and González (2016) for a discussion about assessing equating transformations.

2.6.1 Standard Errors of Equating

Equating error at score x_i for a given equating is (Kolen and Brennan, 2014, pp.248-249)

$$\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)], \tag{2.146}$$

where $\hat{eq}_Y(x_i)$ is an estimate of the Form Y equivalent of a Form X score in the sample and $E[\hat{eq}_Y(x_i)]$ is the expected equivalent, where E is the expectation over random samples from the population(s).
The equating error variance at score point x_i is

$$var[\hat{eq}_Y(x_i)] = E\{\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)]\}^2, \tag{2.147}$$

where the variance is taken over replications. The standard error of equating is

$$se[\hat{eq}_Y(x_i)] = \sqrt{E\{\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)]\}^2}. \tag{2.148}$$

To estimate standard error of equating, we can use the Bootstrap or the Delta method (Kolen and Brennan, 2014, pp.250-276).

2.6.2 The Bootstrap

The bootstrap method (Efron and Tibshirani, 1993) is a method for estimating standard errors of a wide variety of statistics that is computationally intensive. The steps in estimating standard errors of a statistic using the bootstrap from a single sample are as follows:

1. Begin with a sample of size N.

2. Draw a random sample, with replacement, of size N from this sample data. Refer to this sample as a bootstrap sample.
3. Calculate the statistic of interest for the bootstrap sample.
4. Repeat steps 2 and 3 R times.
5. Calculate the standard deviation of the statistic of interest over the R bootstrap samples. This standard deviation is the estimated bootstrap standard error of the statistic.

2.6.3 The Delta Method

The delta method is based on a Taylor series expansion (Kendall and Stuart, 1977). Define for the population $eq_Y(x_i; \Theta_1, \Theta_2, ..., \Theta_t)$ as an equating function of test score x_i and parameters $\Theta_1, \Theta_2, ..., \Theta_t$. By the delta method, an approximate expression for the sampling variance is

$$var[\hat{eq}_Y(x_i)] \cong \sum_j eq'^2_{Y_j} var(\hat{\Theta}_j) + \sum\sum eq'_{Y_j} eq'_{Y_k} cov(\hat{\Theta}_j, \hat{\Theta}_k). \tag{2.149}$$

In this equation, $\hat{\Theta}_j$ is a sample estimate of Θ_j and eq'_{Y_j} is the partial derivative of eq_{Y_j} with respect to Θ_j and evaluated at $x_i, \Theta_1, \Theta_2, ..., \Theta_t$.
In matrix form, if a vector of parameter estimates $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with variance matrix $\hat{\boldsymbol{I}}$, a vector-valued continuously differentiable function f of $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed as well, and its variance is obtained by pre- and postmultiplying $\hat{\boldsymbol{I}}$ with the Jacobian matrix $\boldsymbol{J_f}$ of the function evaluated at the parameter estimates,

$$COV(f(\hat{\boldsymbol{\theta}})) = \boldsymbol{J_f}(\hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}(\boldsymbol{J_f}(\hat{\boldsymbol{\theta}}))^t. \tag{2.150}$$

Equating error variance expressions obtained using the delta method for a few methods and designs are given below as examples (Kolen and Brennan, 2014, pp.261-263).

Linear equating

Holland and Rubin (1982a) presented the partial derivatives and standard errors and covariances between the moments to apply the delta method for linear equating with the equivalent groups design. They showed that

$$\begin{aligned} var[\hat{l}_Y(x_i)] \cong \sigma^2(Y)\Bigg\{ \frac{1}{N_X} + \frac{1}{N_Y} + \left[\frac{sk(X)}{N_X} + \frac{sk(Y)}{N_Y}\right]\left[\frac{x_i - \mu(X)}{\sigma(X)}\right] \\ + \left[\frac{ku(X)-1}{4N_X} + \frac{ku(Y)-1}{4N_Y}\right]\left[\frac{x_i - \mu(X)}{\sigma(X)}\right]^2 \Bigg\}, \end{aligned} \tag{2.151}$$

where $sk(X)$, $sk(Y)$, $ku(X)$ and $ku(Y)$ are the skewness and kurtosis of X and Y.

Thorndike (1971) applied the delta method for linear equating with the NEAT design. He showed that

$$var[\hat{l}_Y(x_i)] \cong \frac{\sigma^2(Y)[1-\rho^2(X,V)]}{N_{tot}}\left\{2+[1+\rho^2(X,V)]\left[\frac{x_i-\mu(X)}{\sigma(X)}\right]^2\right\}. \tag{2.152}$$

In this equation, $\rho(X,V)$ is the correlation between common items and total score, and N_{tot} is the total number of examinees taking the forms.

Equipercentile equating

Lord (1982) used the delta method to develop the following expression for the standard error of equipercentile equating under the equivalent groups design:

$$\begin{aligned} var[\hat{e}_Y(x_i)] \cong \frac{1}{[G(y_U^*)-G(y_U^*-1)]^2}&\left\{\frac{[P(x_i)/100][1-P(x_i)/100](N_X+N_Y)}{N_XN_Y}\right. \\ -\ &\left.\frac{[G(y_U^*)-P(x_i)/100][P(x_i)/100-G(y_U^*-1)]}{N_Y[G(y_U^*)-G(y_U^*-1)]}\right\}. \end{aligned} \tag{2.153}$$

To estimate the error variances, sample values can be substituted in place of the parameters in equation 2.153. The error variance depends on the proportion of examinees at scores on Form Y, as symbolized by $G(y_U^*)-G(y_U^*-1)$. If this quantity were 0, then the error variance would be undefined because of a 0 term in the denominator.

2.6.4 Standard Errors for Scale Scores

A variation of the delta method can be used to estimate the scale score standard errors. To develop this variation, consider a situation in which a parameter Θ is being estimated, where the estimate is symbolized by $\hat{\Theta}$. Also assume that the error variance in estimating the parameter is known, which is symbolized by $var(\hat{\Theta})$. Finally, assume that the estimate is to be transformed using the function f. Approximately, we have: (Kendall and Stuart, 1977)

$$var[f(\hat{\Theta})] \cong f'^2(\Theta)var(\hat{\Theta}), \tag{2.154}$$

where f' is the first derivative of f.
This formulation can be applied to equating by substituting $eq_Y(x_i)$ for the parameter Θ, $\hat{eq}_Y(x_i)$ for $\hat{\Theta}$, and Form Y raw-to-scale score transformation s for the function f. To apply this equation directly, the first derivative of the Form Y raw-to-scale transformation is needed at $eq_Y(x_i)$.

2.6.5 Estimating sample size

In addition to comparing equating error associated with different designs and methods, standard errors of equating also can be useful in specifying the sample size of the examinees required to achieve a given level of equating precision for a particular equating design and method. We give below as examples sample sizes for a few methods and designs.

Equivalent Groups Linear Equating

Suppose that linear equating with the equivalent groups design is to be used. Let u refer to the maximum proportion of standard deviation units that is judged to be appropriate for the standard error of equating. The value of N_{tot} is found that gives a specified value for $u\sigma(Y)$ for the standard error of equating. We have

$$N_{tot} \cong \frac{2}{u^2}\left\{2 + \left[\frac{x_i - \mu(X)}{\sigma(X)}\right]^2\right\}, \tag{2.155}$$

which represents the total sample size required for the standard error of equating to be equal to u standard deviation units on the old form.

Equivalent Groups Equipercentile Equating

In equipercentile equating with the equivalent groups design N_{tot} is given by

$$N_{tot} \cong \frac{4[P(x_i)/100][1 - P(x_i)/100]}{u^2\phi^2}. \tag{2.156}$$

Recall that this equation assumes that the scores on Form X are normally distributed.

2.6.6 The Standard Error for Kernel Equating

Applying the delta method (see subsection 2.6.1) to the kernel method of equating (von Davier, 2011, pp.67-85), $\hat{\boldsymbol{\theta}}$ in equation 2.150 contains the parameters of the statistical model for the score distributions, and the equation function eq plays the role of function f. Hence,

$$COV(\boldsymbol{eq(\hat{\theta})}) = \boldsymbol{J_{eq}}(\hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}(\boldsymbol{J_{eq}}(\hat{\boldsymbol{\theta}}))^t. \tag{2.157}$$

Since eq is a composition of functions, its Jacobian may be computed as the product of their Jacobians:

$$\begin{aligned}\boldsymbol{J_{eq}} &= \frac{deq(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \\ &= \frac{de_X(\boldsymbol{r})}{d\boldsymbol{r}}\frac{dDF(\boldsymbol{p})}{d\boldsymbol{p}}\frac{dg(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \\ &= \boldsymbol{J_e J_{DF} J_g}.\end{aligned} \tag{2.158}$$

2.6.7 Assess the Difference Between Equating Functions

We know that often more than one equating method is applied to the data stemming from a particular test administration. If differences in estimated equating functions are observed, the question arises as to whether these differences reflect real differences in the underlying true equating functions or merely reflect sampling error.
von Davier et al. (2004) presented expressions for the standard error of the difference between two equating functions evaluated in the same score of Y. The equating difference function mapping each score of Y into the difference of equated scores is a vector-valued function with as the jth component

$$\boldsymbol{\Delta^j_{eq}} = (e^j_{1X} - e^j_{2X}) \circ DF \circ g, \tag{2.159}$$

The asymptotic variance matrix is obtained as

$$COV(\boldsymbol{\Delta_{eq}(\hat{\theta})}) = \boldsymbol{J_{\Delta_{eq}}}(\boldsymbol{\hat{\theta}})\boldsymbol{\Sigma}(\boldsymbol{J_{\Delta_{eq}}}(\boldsymbol{\hat{\theta}}))^t, \tag{2.160}$$

where

$$\begin{aligned}\boldsymbol{J_{\Delta_{eq}}} &= \frac{d\Delta_{eq}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \\ &= \frac{d(e_{1X} - e_{2X})(\boldsymbol{r})}{d\boldsymbol{r}}\frac{dDF(\boldsymbol{R})}{d\boldsymbol{R}}\frac{dg(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \\ &= (\boldsymbol{J_{e_1}} - \boldsymbol{J_{e_2}})\boldsymbol{J_{DF} J_g}.\end{aligned} \tag{2.161}$$

von Davier (2011) present a generalization of the Wald test (Wald, 1943) which tests the null hypothesis that there is no difference between the two equating functions. In its general form, the Wald statistic to test a set of linear hypotheses $\boldsymbol{L\beta} = \boldsymbol{0}$,where each row of $\boldsymbol{L}$ represents a linear hypothesis on $\boldsymbol{\beta}$, has the following form:

$$w = (\boldsymbol{L\hat{\beta}})'(\boldsymbol{L}COV(\boldsymbol{\hat{\beta}})\boldsymbol{L'})^{-1}(\boldsymbol{L\hat{\beta}}). \tag{2.162}$$

If $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed, w is asymptotically chi-squared distributed with as degrees of freedom the number of rows of $\boldsymbol{L}$. In the context of testing for a difference between two equating methods, $\boldsymbol{\Delta_{eq}(\hat{\theta})}$ fulfills the role of $\hat{\boldsymbol{\beta}}$, and its covariance matrix is given in Equations 2.160 and 2.161. Hence,

$$w_{\Delta_{eq}} = (\boldsymbol{L\Delta_{eq}(\hat{\theta})})'(\boldsymbol{L}COV(\boldsymbol{\Delta_{eq}(\hat{\theta})})\boldsymbol{L'})^{-1}(\boldsymbol{L\Delta_{eq}(\hat{\theta})}). \tag{2.163}$$

2.7 Some recent ideas in test equating

In this section we shortly describe the idea of local equating and a few local equating methods, a local equating method of cognitively diagnostic modeled observed score, the continuized log-linear method (which is a variant of the kernel method) and a Bayesian nonparametric model for test equating.

2.7.1 Local Observed-Score Equating

The main theorem of local equating follows directly from the equity citerion (see subsection 1.1.6). The equity criterion could be expressed as an equality of the conditional distribution functions $F_{\varphi(Y)|\theta}$ for the equated scores on test Y and $F_{X|\theta}$ for the observed score on test X, that is (von Davier, 2011, pp.201-223):

$$F_{X|\theta}(x) = F_{\varphi(Y)|\theta}(\varphi(y)), \quad \theta \in \mathbb{R}. \tag{2.164}$$

Solving for x by taking the inverse of $F_{X|\theta}$,

$$x = \varphi^*(y;\theta) = F^{-1}_{X|\theta}F_{\varphi(Y)|\theta}(\varphi(y)), \quad \theta \in \mathbb{R}. \tag{2.165}$$

However, because $\varphi(\cdot)$ is monotone, $F_{\varphi(Y)|\theta}(\varphi(y)) = F_{Y|\theta}(y)$. Substitution results in

$$\varphi^*(y;\theta) = F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \quad \theta \in \mathbb{R}, \tag{2.166}$$

as the family of true equating transformations. Equation 2.166 involves the same type of transformation as for the equipercentile equating in $\varphi(y) = F^{-1}(G(y))$, but is now applied to each of the conditional distributions of $X \mid \theta$ and $Y \mid \theta$ instead of only once to the marginal distribution of X and Y for a population of examinees. The fact that the derivation leads to an entire family of transformations reveals a restrictive implicit assumption of our traditional thinking about equating: namely, that equating should be based on a single transformation for the entire population of

choice.
The name *local equating* is derived from the attempt to get as close as possible to the true equating transformations in equation 2.166 to perform the equating. We now describe the basic ideas of a few local equating methods.

Anchor Score as a Proxy of Ability

For a NEAT design it seems natural to use the extra information provided by the anchor test to approximate the true family of equating transformations in Equation 2.166 (Wiberg and van der Linden, 2011). For simplicity, we assume an anchor test A with score A that is not part of Y ("external anchor"). For an anchor test to be usable, A has to be a measure of the same θ as X and Y. Formally, this means a classical true score τ_A that is a monotonic increasing function of the same ability θ as the true scores for X and Y. It should thus hold that $\tau_A = g(\theta)$ where g is an (unknown) monotonically increasing function and θ is the same ability for X and Y. For the distribution of X given θ it holds that

$$f(x \mid \theta) = f(x \mid g^{-1}(\tau_A) = f(x \mid \tau_A), \tag{2.167}$$

Similarly, $f(y \mid \theta) = f(y \mid \tau_A)$. This fact suggests that instead of using an estimate of θ for each examinee, we could use an estimate of τ_A and get the same equating. An obvious estimate of τ_A is the observed score A. The result is an approximation of the family of true transformations in equation 2.166 by

$$\varphi(y; a) = F^{-1}_{X|a}(F_{Y|a}(y)), \quad a = 0, \dots, m, \tag{2.168}$$

where m is the length of the anchor test and $F_{X|a}(x)$ and $F_{Y|a}(y)$ are the distribution functions of X and Y given $A = a$.

$Y = y$ as a Proxy of Ability

The argument for the use of anchor score A as a proxy for θ in the previous method holds equally well for the realized observed score $Y = y$ (Wiberg and van der Linden, 2011). The score can be assumed to have a true score η that is a function of the same ability θ as the true scores on form X. For this reason, we can focus on the distributions of X and Y given η instead of θ. As $Y = y$ is an obvious estimate of η, we have:

$$\varphi(y) = F^{-1}_{X|y}(F_{Y|y}(y)), \quad y = 0, \dots, n. \tag{2.169}$$

In an equating study with a single-group design, the distributions of Y given y are only observable for replicated administrations of form Y to the same examinees. One case for which replications are unnecessary is linear equating conditional on $Y = y$. For this case, we have:

$$x = \varphi(y) = \mu_{X|y} + \frac{\sigma_{X|y}}{\sigma_{Y|y}}(y - \mu_{Y|y}), \quad y = 0, \ldots, n, \tag{2.170}$$

As classical test theory implies $\mu_{Y|y} = y$, we obtain:

$$x = \varphi(y) = \mu_{X|y}, \quad y = 0, \ldots, n. \tag{2.171}$$

For all examinees with score $Y = y$, this local method thus equates the observed scores on Y to their conditional means on X.

2.7.2 Local Equating of Cognitively Diagnostic Modeled Observed Scores

Xin and Jiahui (2015) proposed a local equating method of cognitively diagnostic modeled observed score to address the problem of equating without anchor test as well as the satisfaction of equating criteria.

Cognitive Diagnostic Models (CDMs)

CDMs refer to latent class models that employ item response functions (IRFs) to assess examinees mastery or non-mastery on a set of skills called attributes with an explicit Q^{**}-matrix presenting attribute assignment to each item. The Q^{**}-matrices of the two test forms are denoted as Q^{**}_X and Q^{**}_Y, respectively, the I rows of which represent the I items and the P columns of which correspond to the P attributes. In this study, elements in the Q^{**}-matrix are binary: $q_{ip} = 1$ if getting a correct response to item i without guessing requires the mastery of attribute p, and $q_{ip} = 0$ otherwise. Accordingly, the mastery or non-mastery of each attribute represents examinees latent ability levels in the form of binary attribute vectors, denoted as attribute mastery pattern (AMP) here. Let $\boldsymbol{\alpha^*}$ refer to an AMP and Ω as the set of all AMPs. Following DiBello et al. (2007), for the jth examinee we have:

$$\boldsymbol{\alpha_j^*} = (\alpha^*_{1j}, \alpha^*_{2j}, ..., \alpha^*_{Pj}), \tag{2.172}$$

where $\alpha^*_{jp} = 1$ if examinee j has mastered skill p, and $\alpha^*_{jp} = 0$ otherwise.
Following the notations of Leighton et al. (2004), two matrices of order (P, P) (where P is the number of attributes) to describe the relationships among attributes are used: the adjacency matrix (D) and the reachability matrix (R). The D-matrix indicates the direct relationships among attributes: An element of the value of 1 in the position (p', p'') means that attribute p' is prerequisite to attribute p'' and an element of the value of 0 indicates otherwise. The R-matrix represents both the direct and indirect relationships between pairwise attributes. The R matrix can be calculated using $R = (D + I)^{n^{**}}$, where $1 \leq n^{**} \leq P$, D is the adjacency matrix and I is the identity matrix. The p'th row of the R-matrix specifies all the attributes, including the p'th attribute, for which the p'th attribute is a direct or indirect prerequisite (Leighton et al., 2004). The D-matrix and the R-matrix of the two forms are denoted as D_X, D_Y, R_X, and R_Y, respectively.

Two CDMs used in this study are introduced as follows. The probability an examinee correctly responds to an item is determined as follows: $P(X_{ji} = 1 \mid \boldsymbol{\alpha_j^*}) = 1$ if $\boldsymbol{\alpha}_j^*$ contains $1s$ for all skills required for item i, and $P(X_{ji} = 1 \mid \boldsymbol{\alpha_j^*}) = 0$ otherwise. Thus, a computationally simpler form would be:

$$\eta_{ji} = P(X_{ji} = 1 \mid \boldsymbol{\alpha_j^*}) = \prod_{p=1}^{P} \alpha_{jp}^{q_{ip}}, \tag{2.173}$$

in which P is the number of attributes, q_{ip} is the element of the Q^{**}-matrix at (i, p), and α_{jp} is the pth element of the vector $\boldsymbol{\alpha_j}$. The latent response η_{ji} has two values: $\eta_{ji} = 1$ indicates that examinee j masters all the attributes required for item i, and $\eta_{ji} = 0$ indicates that examinee j lacks at least one of the attributes required for item i. The probability of a correct response is also affected by slipping or guessing, characterized by two-item parameters: $g_i^* = P(X_{ji} = 1|\eta_{ji} = 0)$ and $s_i^* = P(X_{ji} = 0|\eta_{ji} = 1)$. The IRF of the model can be written as:

$$P(X_{ji} \mid \boldsymbol{\alpha_j^*}) = (1 - s_i^*)^{\eta_{ji}} g_i^{*(1-\eta_{ji})}. \tag{2.174}$$

The relationship between the attributes and higher order proficiency can be expressed using the latent logistic regression model:

$$P(\boldsymbol{\alpha_j^*} \mid \theta_j) = \prod_{p=1}^{P} \left\{ \frac{exp[1.7\lambda_1(\theta_j - \lambda_{0p})]}{1 + exp[1.7\lambda_1(\theta_j - \lambda_{0p})]} \right\}, \tag{2.175}$$

where λ_1 and λ_{0p} are the latent discrimination and difficulty parameters, respectively. Attributes with higher λ_{0p} are deemed more difficult to master.

The Criteria of Equating connected to Cognitive Diagnostic Models

The same construct criterion. It is common sense that tests measuring different constructs should never be equated (Kolen and Brennan, 2014). Xin and Jiahui (2015) substained that the necessary and sufficient condition of the same construct criterion for this kind of models would be that the two test forms have (a) the same Q^{**}-matrix or (b) sufficient Q^{**}-matrices (Q^{**}-matrices are sufficient when they include the R-matrix as their submatrix) that are similar to each other.
Equity criterion (see subsection 2.7.1). The observed-score distribution for Form Y of a certain examinee whose AMP is $\boldsymbol{\alpha^*}$ is denoted as $F_{Y|\alpha^*}(y)$, and his or her distribution of the equated observed scores from Form X to the scale of Form Y is denoted as $F_{\varphi(X)|\alpha^*}(\varphi(x))$. Therefore, the equity criterion for the full distributions of observed scores on X and Y given AMP can be written as

$$F_{Y|\alpha^*}(y) = F_{\varphi(X)|\alpha^*}(\varphi(x)), \quad \boldsymbol{\alpha^*} \in \boldsymbol{\Omega}. \tag{2.176}$$

The equity criterion can also be expressed as the requirement of zero equating error for all AMPs:

$$e(x;\boldsymbol{\alpha^*}) = F_{Y|\alpha^*}(y) - F_{\varphi(X)|\alpha^*}(\varphi(x)), \quad \boldsymbol{\alpha^*} \in \boldsymbol{\Omega}. \tag{2.177}$$

The Family of True Transformations for Cognitive Diagnostic Models

When the error for each attribute pattern in Equation 2.175 is required to be 0, y can be solved by taking the inverse of $F_{Y|\alpha^*}$:

$$y = \varphi_{\boldsymbol{\alpha^*}}(x) = F^{-1}_{Y|\alpha^*} F_{\varphi(X)|\alpha^*}(\varphi(x)), \quad \boldsymbol{\alpha^*} \in \boldsymbol{\Omega}. \tag{2.178}$$

Because the function $\varphi(\cdot)$ should be monotone, $F_{\varphi(X)|\boldsymbol{\alpha^*}}(\varphi(x)) = F_{X|\boldsymbol{\alpha^*}}(x)$. Then, the following equation is obtained

$$\varphi_{\boldsymbol{\alpha^*}}(x) = F^{-1}_{Y|\boldsymbol{\alpha^*}} F_{X|\boldsymbol{\alpha^*}}(x), \quad \boldsymbol{\alpha^*} \in \boldsymbol{\Omega}, \tag{2.179}$$

as the family of true equating transformation.

A CDM-Based Local Equating Method

Before equating, a CDM is applied to examinee responses to obtain estimates for item parameters and AMPs. The item parameter estimates are used to calculate the conditional distributions of observed scores given AMP using an adaptation of the recursive algorithm of Lord and Wingersky (1984). The probabilities and distributions are conditional on discrete $\boldsymbol{\alpha^*}$ instead of continuous θ, then the family of transformations in Equation 2.179 can be obtained. Examinees taking either test form have their AMP estimates, so the equating transformation for a given examinee is chosen from the family of estimated equating transformations according to the estimated AMP denoted as $\hat{\boldsymbol{\alpha^*}}$. Upon substitution of $\hat{\boldsymbol{\alpha^*}}$ into Equation 2.179, the following estimated true transformation is obtained:

$$\varphi_{\hat{\boldsymbol{\alpha^*}}}(x) = F^{-1}_{Y|\hat{\boldsymbol{\alpha^*}}} F_{X|\hat{\boldsymbol{\alpha^*}}}(x), \quad \hat{\boldsymbol{\alpha^*}} \in \boldsymbol{\Omega}. \tag{2.180}$$

2.7.3 The Continuized Log-Linear Method

We remember that the Kernel Method has five step:(a) presmoothing, (b) estimating score probabilities, (c) continuization, (d) equating, and (e) calculating the standard error of equating. The *continuized log-linear (CLL) method* directly takes the log-linear function in the presmoothing step and transforms it into a continuous distribution (Wang, 2008). As example, we shortly describe this method for equivalent groups designs.

The CLL method for the Equivalent Groups Design

For the equivalent groups design, the CLL distribution utilizes the polynomial log-linear function obtained in the log-linear smoothing step. The probability density function (PDF) is expressed as (von Davier, 2011, pp.142-143):

$$f(x) = \frac{1}{D^{**}} exp(\boldsymbol{b^T}\boldsymbol{\beta}), \tag{2.181}$$

where $\boldsymbol{b^T} = (1, x, x^2, ..., x^M)$ is a vector of polynomial terms of test X score x, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_M)^T$ is the vector of the parameters, and M is the degree of the polynomial. D^{**} is a normalizing costant that ensures that $f(x)$ is a PDF.
With the CLL continuization method, the design function must be applied after the continuization step, but, for the equivalent groups design, the design function is an identity function, which means that no such procedure is needed.

2.7.4 A Bayesian Nonparametric Model for Test Equating

A Bayesian nonparametric model is a model on an infinite-dimensional parameter space (von Davier, 2011, pp.175-184). The parameter space is typically chosen as the set of all possible solutions for a given learning problem. For example, in a regression problem the parameter space can be the set of continuous functions, and in a density estimation problem the space can consist of all densities (Orbanz and Teh, 2010).

For modeling continuous test score distributions (F_X, F_Y), it is possible to construct a bivariate random Bernstein polynomial prior distribution on (F_X, F_Y) via the random distributions: (Karabatsos and Walker, 2009)

$$F_X(\cdot; G_X, p_X) = \sum_{k=1}^{p_X} \left[G_X\left(\frac{k}{p_X}\right) - G_X\left(\frac{k-1}{p_X}\right) \right] Beta(\cdot \mid k, p_X - k + 1), \tag{2.182}$$

$$F_Y(\cdot; G_Y, p_Y) = \sum_{k=1}^{p_Y} \left[G_Y\left(\frac{k}{p_Y}\right) - G_Y\left(\frac{k-1}{p_Y}\right) \right] Beta(\cdot \mid k, p_Y - k + 1), \tag{2.183}$$

with (G_X, G_Y) coming from the bivariate Dirichlet Process model (Walker and Muliere, 2003), and with independent prior distributions $\pi(p_X)$ and $\pi(p_Y)$ for p_X and p_Y. Given samples of observed scores on the two tests $x_{n(X)} = \{x_1, ..., x_{n(X)}\}$ and $y_{n(Y)} = \{y_1, ..., y_{n(Y)}\}$, the random bivariate Bernstein polynomial prior combines with these data to define a joint posterior distribution, which we denote by $F_X, F_Y \mid x_{n(X)}, y_{n(Y)}$. The Gibbs sampling algorithm can be used to infer the posterior distribution $F_X, F_Y \mid x_{n(X)}, y_{n(Y)}$. At each iteration of this Gibbs algorithm, a current set of p_X, G_X for test X and p_Y, G_Y for test Y is available, from which it is possible to construct random distribution functions (F_X, F_Y) and the random equating function $e_Y(x) = F_Y^{-1}(F_X(x)) = y$.

Chapter 3

Equating with covariates

In this chapter, after a short introduction, equating methods when covariates are used will be discussed.

3.1 Introduction

Equating with covariates is motivated by an interest in the possibility of using information about the examinees' background to increase the precision of the estimators. Kolen (1990) made some comments on matching in equating. He explained that, because equating tends to work well when the examinee groups are similar to one other, it is necessary to identify similar subgroups to obtain more accurate equating and, operationally, matching on the anchor test scores is the easiest way to form similar groups. Livingston et al. (1990) underlined that, ideally, the right variable to use as a basis for matched sampling would be a measure of whatever is causing the old-form and new-form populations to differ systematically in ability. In practice, because we cannot measure this variable, we can select on a "propensity score" (Rosenbaum and Rubin, 1983): it is defined as the linear combination of all the variables we can measure that best discriminates between the old-form and the new-form populations. Cook et al. (1990) discussed the equating of reasonably parallel forms of College Board Achievement Tests in Biology, Chemistry, Mathematics Level II, American History and Social Studies, and French. They explored these questions: what would be the effect on reported scores of changing the currently used sampling plan? And,given that a serious effect was observed, could this effect be ameliorated by matching new- and old-form samples on ability level using distributions of scores on a set of common items or on some other covariate, such as responses to a background questionnaire given at the time of testing or when an examinee first registers to take a test? Their results showed that the equating models were seriously affected by differences in group ability and attempts to ameliorate the effect of group differences by matching on a covariate such as the common-item set itself were quite good. Wright and Dorans (1993) defined the "selection variable" as the variable or set of variables along which subpopulations differ and they discussed whether we can improve equating results if we use the selection variable as either an anchor in an anchor test design or as a variable on which to match equating samples. Results showed that matching on the selection variable improved accuracy over matching on the equating test for all methods examined (Tucker and Levine linear models, chained equipercentile and frequency estimation equipercentile models). Results with the selection variable as an anchor were good for both the Tucker and frequency estimation methods. Liou et al. (2001) examined the possibility of using surrogate variables (e.g., school grades, other test scores, examinee background information) as replacements for common items when common items are unavailable or contaminated by nonrandom errors. Let Test X and Test Y be target tests between which comparable scores must be determined, the authors underlined that the surrogate variable should correlate moderately well with X and Y. Their empirical example showed that the surrogate variable worked as well as the common-item score. Holland et al. (2007) underlined that anchor test scores should be highly correlated with scores on the tests being equated: the importance of this property of the anchor test raises the question of whether it makes sense to use supplemental information about the examinees in both groups to increase the correlation.

3.2 Observed Score Linear Equating with Covariates

Bränberg and Wiberg (2011) suggest a new method for inclusion of background information in equating to improve the precision. Suppose that we have two test forms and one population of potential examinees (in the random groups design) or two populations of potential examinees (in the nonequivalent groups designs). Let Y be the score an examinee would get if tested with test form Y and X the score an examinee would get if tested with test form X. Assume that the following linear model is appropriate in the population (or in both populations, when we have two populations):

$$Y = \boldsymbol{z}^T \boldsymbol{\beta}_Y + \epsilon_Y, \tag{3.1}$$

$$X = \boldsymbol{z}^T \boldsymbol{\beta}_X + \epsilon_X, \tag{3.2}$$

where $\boldsymbol{z}$ is a $kx1$ vector of covariates (including a constant for the intercept), $\boldsymbol{\beta}_Y$ and $\boldsymbol{\beta}_X$ are $kx1$ vectors of coefficients, $\boldsymbol{z}^T\boldsymbol{\beta}_Y$ and $\boldsymbol{z}^T\boldsymbol{\beta}_X$ are the mean test scores in the subpopulation defined by the values on the covariates, and ϵ_Y and ϵ_x reflect the difference between each examinees score and the mean and they have null expected values and variances given by σ_Y^2 and σ_X^2, respectively. The vector of covariates may consist of variables such as gender, age, education, etc. To equate the two test forms, we need an equating function that maps the scores on one test form into equivalent scores on the other test form. The authors assume the existence of a linear equating function

$$Y^* = eq_Y = b_0 + b_1 X, \tag{3.3}$$

where $b_0 = \mu_Y - \mu_X(\sigma_Y \sigma_X^{-1})$, $b_1 = \sigma_Y \sigma_X^{-1}$ and μ_X, μ_Y, σ_X and σ_Y are the observed score population means and population standard deviations of X and Y (Kolen and Brennan, 2014). This linear equating function is assumed to equate the two test forms in every subpopulation defined by the values on the covariates. Note that the equating function is symmetrical. If the two test forms are equated by eq_Y , they also are equated by

$$X^* = eq_X = -\frac{b_0}{b_1} + \frac{1}{b_1}Y. \tag{3.4}$$

Now, suppose that we have two random samples of examinees and that we give one of the test forms to one sample and the other test form to the other sample. Let N_Y and N_X (with $N_Y + N_X = N$) be the number of examinees who take test forms Y and X, respectively. Further, let Y_j and X_j $(j = 1, ..., N_Y, N_{Y+1}, ..., N)$ be the jth examinees (potential) scores on test form Y and test form X, respectively. With random sampling from large populations, the observed test scores can be seen as observations on N (approximately) independent random variables with conditional expectations and variances given by

$$E(Y_j \mid z_j) = \boldsymbol{z}_j^T \boldsymbol{\beta}_Y = E(Y_j^* \mid z_j) = b_0 + b_1 E(X_j \mid z_j), \tag{3.5}$$

$$E(X_j \mid z_j) = \boldsymbol{z}_j^T \boldsymbol{\beta}_X = E(X_j^* \mid z_j) = -\frac{b_0}{b_1} + \frac{1}{b_1} E(Y_j \mid z_j), \tag{3.6}$$

$$V(Y_j \mid z_j) = \sigma_Y^2 = V(Y_j^* \mid z_j) = b_1^2 V(X_j \mid z_j), \tag{3.7}$$

$$V(X_j \mid z_j) = \sigma_X^2 = V(X_j^* \mid z_j) = \frac{1}{b_1^2} V(Y_j \mid z_j). \tag{3.8}$$

The authors obtained the results 3.5-3.8 using equations 3.3 and 3.4 and the definition of equating.

To derive maximum likelihood estimators of the equating parameters, we extend the model with a distributional assumption. Following Potthoff (1966), the authors assume that the conditional distribution of Y_j, given z_j is normal. This assumption implies the following density functions

$$f(y_j \mid \boldsymbol{z}_j) = \frac{1}{\sqrt{\sigma_Y^2 2}} exp\left(-\frac{1}{2\sigma_Y^2} (y_j - \boldsymbol{z}_j^T \boldsymbol{\beta}_Y)^2 \right), \tag{3.9}$$

$$f(x_j \mid \boldsymbol{z}_j) = \frac{b_1}{\sqrt{\sigma_Y^2 2}} exp\left(-\frac{1}{2\sigma_Y^2} (b_0 + b_1 x_j - \boldsymbol{z}_j^T \boldsymbol{\beta}_Y)^2 \right). \tag{3.10}$$

Calculating the log-likelihood function, equating the partial derivatives of the log-likelihood function to zero and finally solving equations for $\hat{b}_0$, $\hat{b}_1$, σ^2_Y and $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}$ we obtain:

$$\hat{b}_0 = \frac{1}{N_X} \sum_{j=N_Y+1}^{N} \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}} - \hat{b}_1 \bar{x}, \tag{3.11}$$

where

$$\bar{x} = \frac{1}{N_X} \sum_{i=N_Y+1}^{N} x_j, \tag{3.12}$$

is the sample mean test score on test form X;

$$\begin{aligned} \hat{b}_1 = & \frac{\sum_{j=N_Y+1}^{N} x_j \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}} - \bar{x} \sum_{j=N_Y+1}^{N} \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}}{2(N_X - 1)s^2_X} \\ & + \sqrt{\left(\frac{\sum_{j=N_Y+1}^{N} x_j \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}} - \bar{x} \sum_{j=N_Y+1}^{N} \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}}{2(N_X - 1)s^2_X} \right) + \frac{N_X \hat{\sigma}^2_Y}{(N_X - 1)s^2_X}}, \end{aligned} \tag{3.13}$$

where

$$s^2_X = \frac{1}{N_X - 1} \sum_{j=N_Y+1}^{N} (x_j - \bar{x})^2, \tag{3.14}$$

is the sample variance for scores on test form X;

$$\sigma^2_Y = \frac{1}{N} \left(\sum_{j=1}^{N_Y} (y_j - \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}})^2 + (\hat{b}_0 + \hat{b}_1 x_j - \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Y}})^2 \right); \tag{3.15}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}} = \left((\sum_{j=1}^{N} \boldsymbol{z}_j \boldsymbol{z}_j^T)^{-1} (\sum_{j=1}^{N_Y} \boldsymbol{z}_j y_j + \hat{b}_1 \sum_{j=N_Y+1}^{N} \boldsymbol{z}_j x_j + \hat{b}_0 \sum_{j=N_Y+1}^{N} \boldsymbol{z}_j) \right). \tag{3.16}$$

The maximum likelihood equations can be solved iteratively using the following procedure:

1. Set $k^* = 0$ and select starting values $\hat{b}_0^{(k^*)}$ and $\hat{b}_1^{(k^*)}$ for $\hat{b}_0$ and $\hat{b}_1$, respectively (in most applications the starting values $\hat{b}_0^{(0)} = 0$ and $\hat{b}_1^{(0)} = 1$ will be adequate).

2. Use $\hat{b}_0^{(k^*)}$ and $\hat{b}_1^{(k^*)}$ in equation 3.16 to get $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{\boldsymbol{(k^*+1)}}$ as an estimate of $\boldsymbol{\beta_Y}$.

3. Use $\hat{b}_0^{(k^*)}$, $\hat{b}_1^{(k^*)}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{\boldsymbol{(k^*+1)}}$ in equation 3.15 to get $\sigma_Y^{2(\hat{k^*}+1)}$ as an estimate of σ_Y^2.

4. Use $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{\boldsymbol{(k^*+1)}}$ and $\sigma_Y^{2(\hat{k^*}+1)}$ in equation 3.13 to get $\hat{b}_1^{(k^*+1)}$.

5. Use $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}}^{\boldsymbol{(k^*+1)}}$ and $\hat{b}_1^{(k^*+1)}$ in equation 3.11 to get $\hat{b}_0^{(k^*+1)}$.

6. Increase k^* by one and repeat steps 2-5 until convergence.

The authors use a simulation study to examine the effect on the estimators of using the model with covariates and they also give a numerical illustration of the use of the model with real data. The simulation study indicates that, with a random groups design, the accuracy of the estimators can be improved by using covariates and it also indicated that, with a nonequivalent groups design, the most efficient way to estimate the equating parameters is to use a good anchor test. The accuracy of the estimators can be improved by supplementing anchor test scores with observations on covariates, but the improvement is small. In the absence of an anchor test, the equating can be improved if covariates are used to correct for differences between the groups, provided that we can find covariates that can explain these differences. The numerical illustration indicates that, using a nonequivalent groups design with no anchor test, but with covariates, results are obtained which are close to the results from the equating actually performed.

3.3 Kernel Equating Under the Non-Equivalent Groups With Covariates Design

Wiberg and Bränberg (2015) describe a method for performing kernel equating with the NEC (non-equivalent groups with covariates) design. The authors assume that they have a sample of examinees from each of two populations P and Q and that for each examinee, they also have observations for other variables correlated with the test scores $\boldsymbol{X}$ and $\boldsymbol{Y}$. The idea is to use information from these covariates as a substitute for an anchor test to control for differences in ability between the two groups. $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_D)^t$ is denoted as the column vector of D discrete covariates. Observations on V are denoted $\boldsymbol{v}_l, l = 1, ..., L$ where L is the number of all possible combinations of values for the covariates $\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_D$. They implemented these ideas in the five steps of kernel equating as described next.

Presmoothing

In presmoothing, a statistical model is fitted to the empirical distribution obtained from the sampled data. The authors use a log-linear model because such a method is commonly used in the kernel method of test equating (von Davier et al., 2004). Let N_{Pjl} and N_{Qkl} be the number of examinees in the sample from populations P and Q, respectively, and observation on $\boldsymbol{X}$ and $\boldsymbol{Y}$ be denoted by $x_j, j = 1, ..., J$ and $y_k, k = 1, ..., K$, respectively. Thus, they have $\boldsymbol{X} = x_j$, $\boldsymbol{Y} = y_k$, and $\boldsymbol{V} = \boldsymbol{v}_l$. The vectors $\boldsymbol{N}_P = (N_{P11}, ..., N_{PJL})^t$ and $\boldsymbol{N}_Q = (N_{Q11}, ..., N_{QKL})^t$ are assumed to be independent and they each have a multinomial distribution. $p_{jl} = P(\boldsymbol{X} = x_j, \boldsymbol{V} = \boldsymbol{v}_l \mid P)$ and $q_{kl} = P(\boldsymbol{Y} = y_k, \boldsymbol{V} = \boldsymbol{v}_l \mid Q)$ are the joint probabilities, and following log-linear models are assumed for p_{jl} in population P and q_{kl} in population Q:

$$logp_{jl} = \alpha_P + \boldsymbol{u}_{Pjl}^t \boldsymbol{\beta}_P, \quad logq_{kl} = \alpha_Q + \boldsymbol{u}_{Qkl}^t \boldsymbol{\beta}_Q, \tag{3.17}$$

where α_P and α_Q are normalizing constants to ensure that the probabilities sum to 1, $\boldsymbol{u}_{Pjl}$ and $\boldsymbol{u}_{Qkl}$ are vectors of known constants, and $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$ are vectors of unknown parameters.

Estimation of the Score Probabilities

The target population T is defined as $T = wP + (1-w)Q, 0 \leq w \leq 1$. Let r_j be the probability of a randomly selected individual in population T scoring x_j on Test X and s_k be the probability on Test Y. The score distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ over T can be obtained from

$$r_j = P(\boldsymbol{X} = x_j \mid T) = wr_{Pj} + (1-w)r_{Qj}, \tag{3.18}$$

$$s_k = P(\boldsymbol{Y} = y_k \mid T) = ws_{Pk} + (1-w)s_{Qk}, \tag{3.19}$$

where $r_{Pj} = P(\boldsymbol{X} = x_j \mid P)$, $r_{Qj} = P(\boldsymbol{X} = x_j \mid Q)$, $s_{Pk} = P(\boldsymbol{Y} = y_k \mid P)$ and $s_{Qk} = P(\boldsymbol{Y} = y_k \mid Q)$ are the score probabilities of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the two populations P and Q. The distribution of $\boldsymbol{X}$ in P and of $\boldsymbol{Y}$ in Q can be estimated through the estimators

$$\hat{r}_{Pj} = \sum_l \hat{p}_{jl}, \quad \hat{s}_{Qk} = \sum_l \hat{q}_{kl}. \tag{3.20}$$

The problem is that there are no data to directly estimate r_{Qj} and s_{Pk}. It is assumed that the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{V}$ and the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{V}$ are the same in both populations. With these two assumptions, the distribution of $\boldsymbol{X}$ in population Q and the distribution of $\boldsymbol{Y}$ in population P can be estimated by the following:

$$\hat{r}_{Qj} = \sum_l \left(\frac{\hat{p}_{jl}}{\sum_j \hat{p}_{jl}} \sum_k \hat{q}_{kl} \right), \quad \hat{s}_{Pk} = \sum_l \left(\frac{\hat{q}_{kl}}{\sum_k \hat{q}_{kl}} \sum_j \hat{p}_{jl} \right). \tag{3.21}$$

Continuization

Because test scores are typically discrete, a Gaussian kernel is used in kernel equating to continuize the two estimated discrete cumulative distribution functions (von Davier et al., 2004). This means that instead of the discrete variable $\boldsymbol{X}$, a continuous variable is used:

$$\boldsymbol{X}(h_X) = a_X(\boldsymbol{X} + h_X\boldsymbol{Z}) + (1 - a_X)\mu_X, \quad where \quad a_X = \sqrt{\frac{\sigma_X^2}{(\sigma_X^2 + h_X^2)}}, \tag{3.22}$$

and where the bandwidth $h_X > 0$ is a constant, $\boldsymbol{Z}$ is a standard normal variable independent of $\boldsymbol{X}$, and $\mu_X = \sum_j x_j r_j$ is the mean of $\boldsymbol{X}$ in population T. The term a_X is a constant where σ_X^2 is the variance of $\boldsymbol{X}$ in population T. The variable $\boldsymbol{X}(h_X)$ has the same mean and the same variance as $\boldsymbol{X}$. The Gaussian kernel smoothing of the distribution of $\boldsymbol{X}$ is given by

$$F_{h_X}(x) = P(\boldsymbol{X}(h_X) \leq x) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right), \tag{3.23}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The Gaussian kernel smoothing of the distribution of $\boldsymbol{Y}$ is given by

$$G_{h_Y}(y) = P(\boldsymbol{Y}(h_Y) \leq y) = \sum_k s_k \Phi\left(\frac{y - a_Y y_k - (1 - a_Y)\mu_Y}{a_Y h_Y}\right). \tag{3.24}$$

The bandwidths in the kernel equating framework are usually selected by minimizing a penalty function (von Davier, 2013), altough other methods exist (Häggström and Wiberg, 2014). To find h_X, one must minimize the penalty function

$$PEN(h_X) = \sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2 + \kappa \sum_j A_j, \tag{3.25}$$

where $\hat{f}_{h_X}$ is the estimated density function of $\boldsymbol{X}(h_X)$ and κ is a constant typically equal to 0 or 1 depending on whether the second penalty term is used. $A_j = 1$ if either of the following conditions is satisfied: (a) $\hat{f}'_{h_X}(x_j - v) > 0$ and $\hat{f}'_{h_X}(x_j + v) < 0$ or (b) $\hat{f}'_{h_X}(x_j - v) < 0$ and $\hat{f}'_{h_X}(x_j + v) > 0$, where $\hat{f}'_{h_X}(x_j)$ is the derivative of $\hat{f}_{h_X}(x_j)$ and v is a constant typically chosen to be 0.25 (von Davier, 2013). Otherwise, $A_j = 0$.

Equating

The estimator of the equipercentile kernel equating function for equating $\boldsymbol{X}$ to $\boldsymbol{Y}$ in population T is obtained from equations 3.23 and 3.24

$$\hat{e}_Y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)). \tag{3.26}$$

Standard Error of Equating

The estimated equated score $\hat{e}_Y(x)$ is based on a sample; thus there is a random error associated with this point estimator. The standard error of equating (SEE) is an estimate of each estimator's standard deviation and, in the kernel equating framework, it is defined as (von Davier et al., 2004)

$$SEE_Y(x) = \| \hat{\boldsymbol{J}}_{e_y} \hat{\boldsymbol{J}}_{DF} \boldsymbol{C} \|, \tag{3.27}$$

where $\| \quad \|$ denotes the Euclidean norm of the vectors and $\boldsymbol{C}$ is a matrix that is based on the covariances between the estimators of the probabilities in $\boldsymbol{P}$ and $\boldsymbol{Q}$. $\hat{\boldsymbol{J}}_{e_y}$ is the Jacobian of the equated score, and $\hat{\boldsymbol{J}}_{DF}$ is the Jacobian of the Design Function, that is, a function that maps probabilities in $\boldsymbol{P}$ and $\boldsymbol{Q}$ into $\boldsymbol{r}$ and $\boldsymbol{s}$ where $\boldsymbol{r} = (r_1, ..., r_j)^t$ and $\boldsymbol{s} = (s_1, ..., s_K)^t$. The authors explain how to obtain the parts needed to calculate the SEE.

$\boldsymbol{P}$ is the JxL matrix of probabilities p_{jl} in population P, and $\boldsymbol{Q}$ is the KxL matrix of probabilities q_{kl} in population Q. Let $v(\boldsymbol{P}) = (\boldsymbol{p}_{11}, ..., \boldsymbol{p}_{JL})^t$ and $v(\boldsymbol{Q}) = (\boldsymbol{q}_{11}, ..., \boldsymbol{q}_{KL})^t$ be vectorized versions of $\boldsymbol{P}$ and $\boldsymbol{Q}$. These probabilities are estimated in the presmoothing step. Let t_{P_m} and t_{Q_m} be the sum of the mth column in matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. The jth element of the vector $\boldsymbol{r}$ is then

$$r_j = wr_j + (1-w)r_{Qj} = \sum_{m=1}^{L} \left[w + \frac{(1-w)t_{Q_m}}{t_{P_m}} \right] p_{jm}, \tag{3.28}$$

and s_k is obtained similarly. Thus, the DF for the NEC design is defined as

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{s} \end{pmatrix} = DF(\boldsymbol{P}, \boldsymbol{Q}; w) = \begin{pmatrix} \sum_m \left[w + \frac{(1-w)t_{Q_m}}{t_{P_m}} \right] \boldsymbol{p}_m \\ \sum_m \left[(1-w) + \frac{wt_{P_m}}{t_{Q_m}} \right] \boldsymbol{q}_m \end{pmatrix}, \tag{3.29}$$

where $\boldsymbol{p}_m$ and $\boldsymbol{q}_m$ are the mth column of $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively, and the summation is over all possible combinations of values of the covariates. The Jacobian of the DF is obtained by using $v(\boldsymbol{P})$ and $v(\boldsymbol{Q})$

$$\boldsymbol{J}_{DF} = \begin{pmatrix} \frac{d\boldsymbol{r}}{dv(\boldsymbol{P})} & \frac{d\boldsymbol{r}}{dv(\boldsymbol{Q})} \\ \frac{d\boldsymbol{s}}{dv(\boldsymbol{P})} & \frac{d\boldsymbol{s}}{dv(\boldsymbol{Q})} \end{pmatrix}. \tag{3.30}$$

Next, for F and G the Jacobian $\hat{\boldsymbol{J}}_{e_y}$ is defined as

$$\hat{\boldsymbol{J}}_{e_y} = \left(\frac{de_Y}{dr}, \frac{de_Y}{ds} \right) = \frac{1}{G'} \left(\frac{dF}{dr}, -\frac{dG}{ds} \right), \tag{3.31}$$

where G' is the density of G evaluated at $e_Y(x)$. The estimate of the Jacobian is obtained by substituting $\hat{F}_{h_X}$ and $\hat{G}_{h_Y}$ for F and G. Finally, to obtain the $\boldsymbol{C}$ matrix the authors assume that $\boldsymbol{P}$ and $\boldsymbol{Q}$ are estimated independently using log-linear models and maximum likelihood. This leads to

$$\begin{aligned} cov \begin{pmatrix} v(\hat{\boldsymbol{P}}) \\ v(\hat{\boldsymbol{Q}}) \end{pmatrix} &= \begin{pmatrix} \boldsymbol{C}_{v(\hat{\boldsymbol{P}})}\boldsymbol{C}^t_{v(\hat{\boldsymbol{P}})} & 0 \\ 0 & \boldsymbol{C}_{v(\hat{\boldsymbol{Q}})}\boldsymbol{C}^t_{v(\hat{\boldsymbol{Q}})} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{C}_{v(\hat{\boldsymbol{P}})} & 0 \\ 0 & \boldsymbol{C}_{v(\hat{\boldsymbol{Q}})} \end{pmatrix} \begin{pmatrix} \boldsymbol{C}^t_{v(\hat{\boldsymbol{P}})} & 0 \\ 0 & \boldsymbol{C}^t_{v(\hat{\boldsymbol{Q}})} \end{pmatrix} \\ &= \boldsymbol{C}\boldsymbol{C}^t, \end{aligned} \tag{3.32}$$

where $\boldsymbol{C}$ is obtained from factorization of the covariance matrix. The size of the covariance between r and s is assumed to be small to simplify the calculation of SEE. Note that if it assumed that $\begin{pmatrix} v(\hat{\boldsymbol{P}}) \\ v(\hat{\boldsymbol{Q}}) \end{pmatrix}$ is normally distributed with mean $\begin{pmatrix} v(\boldsymbol{P}) \\ v(\boldsymbol{Q}) \end{pmatrix}$ and variance $\boldsymbol{C}\boldsymbol{C}^t$, then for each x, the estimator of the equated score, $\hat{e}_Y(x)$, given by equation 3.26 is also normally distributed.

The authors illustrate the NEC design with real data. As a comparison, the results are displayed if a random groups design is used; the results if a regular NEAT design is used either with chain equating or PSE; and the results if a NEC design is used with the covariates and an anchor test (NEATNEC design). The comparison of SEE between the NEC, NEAT Chain, NEAT PSE, and the NEATNEC design exhibits similar (large) SEE values at the lower end of the scale and also at the upper end end of the scale. Moreover the NEC design has a smaller SEE than the random groups design and a larger SEE than the NEAT designs, but, using both covariates and the anchor test (NEATNEC design), they obtain the smallest SEE over a large range of test scores.

3.4 A Dependent Bayesian Nonparametric Model for Test Equating

In this section we shortly describe a Bayesian model for test equating which allows the use of covariates.

Dependent prior probability models

González et al. (2015) proposed a flexible Bayesian nonparametric model for test equating which allows the use of covariates in the estimation of the score distribution functions that lead to the equating transformation.
Let $x_1, ..., x_n$ be observed data defined on a sample space $\mathscr{X}$, and distributed according to a probability distribution F_θ, belonging to a known family $\mathscr{F} = \{F_\theta : \theta \in \Theta\}$ (in the context of equating

θ will be the examinee's ability, here it is a generic parameter). This setup is referred to as a *parametric* model whenever Θ is assumed to be a subset of a finite dimensional space. In the parametric Bayesian framework, a prior $p(\theta)$ is defined on Θ. Parametric Bayesian inference is then based on the posterior distribution $p(\theta \mid x)$, which is proportional to the product of the prior $p(\theta)$ and the likelihood $p(x \mid \theta)$. If $\mathscr{X}$ is an infinite set then $\mathscr{F}$ is infinite-dimensional, and the corresponding prior model $p(F)$ on $\mathscr{F}$ is termed *nonparametric.*

In many applications, it is desirable to allow for dependence of the data on covariates. Under a Bayesian Nonparametric (BNP) approach it is desired to define a probability model that features a set of covariate-dependent continuous distributions $\mathscr{F} = \{F_z : z \in \mathscr{Z}\}$, where now the entire shape of F changes with z, and not just the mean or some other particular functional of the distribution. The nonparametric model is then changed to $x_1, ..., x_n \mid F_z \overset{iid}{\sim} F_z$ with a corresponding prior $p(\mathscr{F})$ on $\mathscr{F} = \{F_z : z \in \mathscr{Z}\}$. Such models are known as *dependent nonparametric models.*

Dependent Dirichlet processes

We remember that the Dirichlet processes (DP) can be constructed as

$$F(\cdot) = \sum_{i=1}^{\infty} \omega_i \delta_{\theta_i}(\cdot), \tag{3.33}$$

where $\delta_{\theta_i}(\cdot)$ denotes a point mass at θ_i, $\theta_i \overset{iid}{\sim} F^*$, $\omega_i = U_i \prod_{j<i}(1-U_j)$, and $U_i \overset{iid}{\sim} Beta(1, M)$. MacEachern (2000) proposed the following dependent Dirichlet process (DDP):

$$F_z(\cdot) = \sum_{i=1}^{\infty} \omega_i(z) \delta_{\theta_i(z)}(\cdot), \tag{3.34}$$

where $\omega_i(z) = U_i(z) \prod_{j<i}(1-U_j(z))$ and both $\theta_i(z)$ and $U_i(z)$ are independent stochastic processes with $U_i \overset{iid}{\sim} Beta(1, M)$.

Dependent Bernstein polynomial priors

We remember that the q-th order Bernstein polynomial of H is defined by

$$B(t; q, H) = \sum_{e=0}^{q} H\left(\frac{e}{q}\right) \binom{q}{e} t^e (1-t)^{q-e}, \tag{3.35}$$

with derivative

$$b(t; q, H) = \sum_{e=1}^{q} w_{eq} \beta(t \mid e, q-e+1), \tag{3.36}$$

where $w_{eq} = H(e/q) - H((e-1)/q)$, and $\beta(\cdot \mid a, b)$ denotes the beta density with parameters a and b.
We can also define a dependent Bernstein polynomial process (DBPP). González et al. (2015) used a simplified version of the general DBPP model which is called single weights DBPP and it is denoted by wDBPP (Barrientos et al., 2012). Consider then $q \sim p(q \mid \lambda)$, where $p(q \mid \lambda)$ is the Poisson(λ) distribution, truncated to $\{1, 2, ...\}$. The model then becomes

$$s(z)(\cdot) = \sum_{e=1}^{\infty} w_e \beta(\cdot \mid \lceil q\theta_e(z) \rceil, q - \lceil q\theta_e(z) \rceil + 1), \tag{3.37}$$

where $\lceil \cdot \rceil$ denotes the ceiling function, $\theta_e(z) = h_z(r_e(z))$, and

$$w_e = v_e \prod_{i<e} [1 - v_i], \tag{3.38}$$

and where $r_1, r_2, ...$ are iid real-valued stochastic processes with probability measure indexed by the parameter Ψ, h_z is a function defined on a set $\mathscr{H} = \{h_z : z \in \mathscr{Z}\}$ of known bijective continuous functions, and $v_1, v_2, ...$ are independent random variables with common distribution indexed by a parameter α. We denote 3.38 by $\mathscr{S} = \{S_z = S(z) : z \in \mathscr{Z}\} \sim wDBPP(\alpha, \lambda, \Psi, \mathscr{H})$.

Dependent BNP model for equating

Let T be the random variable denoting the scores with $S(t)$ the associated probability distribution function. For a vector of covariates z, we are interested in modeling the covariate-dependent score distributions, which will be used to obtain the equating transformation $\varphi(t; z)$. We will denote this as $S_z(t) = S(t \mid Z = z)$. We need to specify a prior probability model for the set $\{S_z : z \in \mathscr{Z}\}$. We use a DBPP so that,

$$\mathscr{S} = \{S_z : z \in \mathscr{Z}\} \sim wDBPP(\alpha, \lambda, \Psi, \mathscr{H}). \tag{3.39}$$

As an example, assume that a new test form X is to be equated to an old form Y. In this case the covariate values denote the test form administered, that is, $Z \in \{X, Y\}$. The c.d.f. of interest would be $S(t \mid Z = z)$. Thus having score data and assuming $\mathscr{S} = \{S_z = S(z) : z \in \mathscr{Z}\} \sim wDBPP(\alpha, \lambda, \Psi, \mathscr{H})$ we can express the equating function as

$$\varphi(t; z) = S^{-1}(S(t \mid Z = X \mid Z = Y)), \tag{3.40}$$

where in this case $\varphi(t; z)$ corresponds to the transformation function which puts scores on version X in the scale of version Y.

Chapter 4

IRT Equating with Covariates: a proposal

A possible way to use covariates in IRT equating might be to include them in the calculation of probability $P(y_{ji})$, where y_{ji} denote the response of examinee j to item i, and subsequently using this probability to conduct IRT equating. The questions thus become: how to include covariates in the calculation of the probability? How to use the probability estimated using covariates to conduct IRT equating? Regarding the first question, Tay et al. (2011) proposed the MM-IRT-C (Mixed-Measurement Item Response Theory With Covariates) model, which at the same time models covariates and latent classes of respondents. The graphical conceptualization of this model is shown in Figure 4.1.

Regarding the second question, we propose two new methods that complement the MM-IRT-C model within the IRT equating.

The structure of this chapter is as follows. First, we describe the MM-IRT-C model and its precursor models. Second, we describe the two new methods that let us to conduct IRT equating using the MM-IRT-C model. Third, a simulation study is used to examine the properties of the new methods. Fourth, a real data example is given to show how the new methods can be used in practice.

4.1 Mixed-Measurement Item Response Theory With Covariates (MM-IRT-C) model and its Precursor Models

To use the IRT-C Model is a possible approach to overcome the limitations of traditional IRT-Differential item functioning (DIF) procedures, because we can assess DIF across multiple variables simultaneously.

Traditional IRT-DIF procedures. DIF occurs when people from different groups (commonly gender or ethnicity) with the same latent trait (ability/skill) have a different probability of giving a certain response on a questionnaire or test (Embretson and Reise, 2000). Uniform DIF is the simplest type of DIF where item characteristic curves (ICC) are equally discriminating, but they exhibit differences in the difficulty parameters. Nonuniform DIF implies that an item can simultaneously vary in discrimination for the different groups while also varying in difficulty. The most famous traditional IRT-DIF procedures were proposed by Lord (1980), Raju (1988) and Swaminathan and Rogers (1990) and are described here.

Lord's Wald test (Lord, 1980) for DIF detection compares vectors of IRT item parameters between groups. For a given item, if the vectors of its parameters differ significantly between groups, then the trace lines differ across groups, and thus the item is functioning differentially for the groups studied.

Raju's Area Measures estimate the area between two item response functions (IRFs) to provide useful information regarding the presence of DIF. Under IRT, if this area is nonzero, the item is functioning differentially in the two groups. Raju (1988) offers formulas for computing the exact area between two ICCs for the one-, two-, and three-parameter logistic IRT models (1PLM, 2PLM and 3PLM, respectively).

The logistic regression model can be used to model differential item functioning by specifying separate equations for the two groups of interest (Swaminathan and Rogers, 1990):

$$P(u_{jj^*} \mid \theta_{jj^*}) = \frac{e^{(\beta_{0j^*}+\beta_{1j^*}\theta_{1j^*})}}{1+e^{(\beta_{0j^*}+\beta_{1j^*}\theta_{1j^*})}}, \quad j = 1, ..., N_{j^*}, j^* = 1, 2. \tag{4.1}$$

Here u_{jj^*} is the response of examinee j in group j^* to the item, β_{0j^*} is the intercept parameter for group j^*, β_{1j^*} is the slope parameter for group j^*, and θ_{jj^*} is the ability of examinee j in group j^*. An alternative but equivalent formulation of model 4.1 is

$$P(u = 1) = \frac{e^z}{[1 + e^z]}, \tag{4.2}$$

where $z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$. In this latter formulation, the variable g represents group membership, where $g = 1$ if examinee belongs to Group 1 and $g = 0$ if examinee belongs to Group 2. The term θg is the product of the two independent variables, g and θ; τ_0 is the intercept parameter, τ_1 corresponds to the ability difference in performance on the item, τ_2 corresponds to the group difference in performance on the item, and τ_3 corresponds to the interaction between group and ability. An item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$ and nonuniform DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$). The hypotheses of interest are therefore that $\tau_2 = 0$ and $\tau_3 = 0$. Swaminathan and Rogers (1990) showed that the statistic for testing this hypothesis has an asymptotic $\boldsymbol{\chi}^{\mathbf{2}}(2)$ distribution.

The IRT-C Model

The traditional IRT-DIF procedures of Lord (1980) and Raju (1988) presented above are limited because (a) multiple observed characteristics cannot be tested for DIF simultaneously, (b) continuous observed characteristics cannot be directly used, and (c) testing of DIF in multiple groups

(> 2) is often engaged in a "piece-meal" fashion.

The IRT-C Model was proposed by Tay et al. (2011) in contrast to traditional DIF detection strategies. This is a single-class/restricted MM-IRT-C approach, and it is a procedure akin to the logistic regression (LR) method for testing observed DIF (Swaminathan and Rogers, 1990) but uses the latent trait score θ_j (as shown in Equation 4.3) rather than the observed total score Y_j. This model is graphically depicted in panel 1A in Figure 4.1, where differences in item discrimination and item location can be tested via Paths 2 and 3, respectively; thus, we can determine whether there are significant differences in the item discriminations (corresponding to nonuniform DIF) and item locations (corresponding to uniform DIF) among observed groups. Furthermore, because a vector of observed characteristics z_j (either continuous or nominal) can be used, it overcomes the three specific limitations of traditional IRT-DIF procedures described above.

2PLM is utilized to describe the relationship between the probability of item endorsement and the latent trait level θ_j. Let y_{ji} denote the response of examinee $j(j = 1, ..., J)$, on item $i(i = 1, ..., I)$. The probability of item endorsement is then

$$P(y_{ji} \mid \theta_j) = \frac{1}{1 + exp(-[a_i\theta_j + b_i])}, \tag{4.3}$$

where a_i and b_i represent the item discrimination and item location, respectively. DIF occurs when the expected score given the same latent trait θ_j is different by virtue of observed characteristic (z_j) (Hulin et al., 1983). DIF can be represented as

$$P(y_{ji} \mid \theta_j, z_j) = \frac{1}{1 + exp(-[a_i\theta_j + b_i + c_i z_j + d_i z_j \theta_j])}, \tag{4.4}$$

where the probability of item responding depends not only on θ_j but also on z_j. The additional terms in equation 4.4, $c_i z_j$ and $d_i z_j \theta_j$, represent the direct and interaction effects for modeling uniform and non-uniform DIF, respectively.

The MM-IRT model

The MM-IRT model (see panel 1B in Figure 4.1) may be viewed as an extension of the IRT model, where latent classes (k) of examinees underlie the set of observed responses. Thus, the conditional probabilities of responding to each item become $P(y_{ji} \mid k, \theta_j)$; which depends not only on the latent trait standing (θ_j) but also on the latent class (k),

$$P(y_{ji} \mid k, \theta_j) = \frac{1}{1 + exp(-[a_{ik}\theta_j + b_{ik}])}. \tag{4.5}$$

Let $\underset{\sim}{\mathrm{y_j}}$ be the vector containing all I item responses ($i = 1, ..., I$) for candidate j; the MM-IRT model is then

$$P(\underset{\sim}{\mathrm{y_j}}) = \sum_{k=1}^{K} \pi_k \int \prod_{i=1}^{I} P(y_{ji} \mid k, \theta_j) f(\theta_j) d\theta_j, \tag{4.6}$$

where the unconditional class membership probabilities π_k serve as weights and sum to 1,$\sum_{k=1}^{K} \pi_k = 1$; $f(\theta_j)$ is taken as the standard normal density and $d\theta_j$ represents the latent trait over which the integration is performed. Similar to Figure 1A, item parameters may have distinct item discriminations or locations across the unobserved groups (or latent classes), indicating unobserved DIF. Unobserved DIF can be tested across LCs in the MM-IRT framework via Paths 5 and 6 in Figure 4.1 panel 1B; significant effects indicate significantly different item discriminations and item locations, respectively.

The MM-IRT-C model

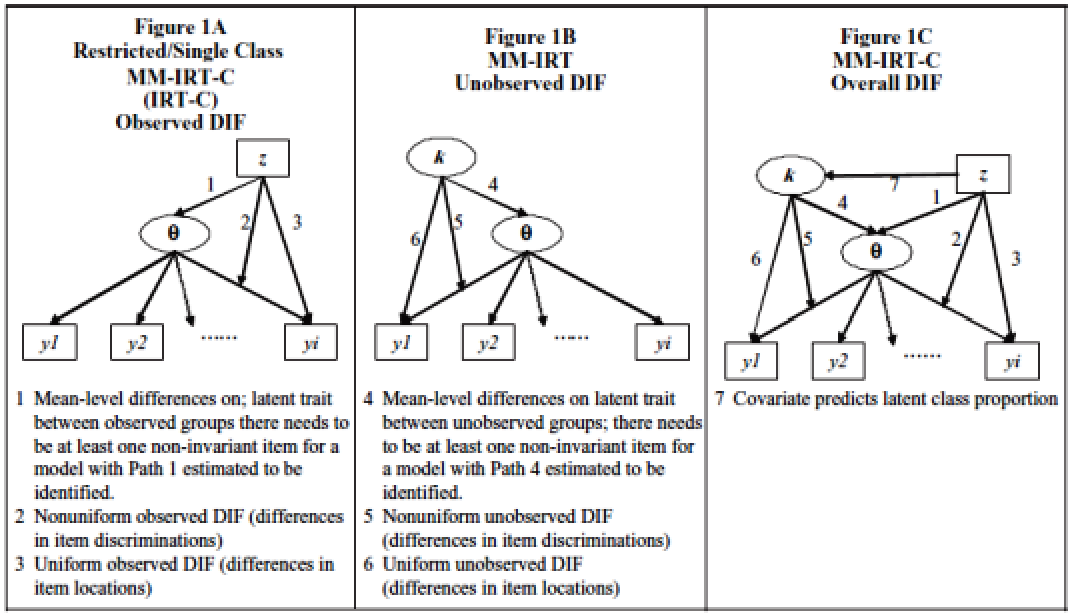
Instead of using a two-step procedure where MM-IRT LCs are first obtained and then associated with other external variables (Eid and Rauber, 2000), we can model the associations of LCs with external observed characteristics within a single, integrated model. Figure 4.1 panel 1C presents the graphical depiction of the MM-IRT with a covariate, or observed characteristic, z. From the model, we see that the association between inferred latent classes and an external covariate (z) is modeled by Path 7; that is, observed group membership may predict latent class membership. For multiple covariates ($p = 1, ..., P$), the MM-IRT-C model may be written as

$$P(\underset{\sim}{\mathrm{y_j}} \mid \underset{\sim}{\mathrm{z_j}}) = \sum_{k=1}^{K} \pi_{k|\underset{\sim}{\mathrm{z_j}}} \int \prod_{i=1}^{I} P(y_{ji} \mid k, \theta_j) f(\theta_j \mid \underset{\sim}{\mathrm{z_j}}) d\theta_j, \tag{4.7}$$

where $\underset{\sim}{\mathrm{z_j}}$ is the covariate vector (nominal or continuous) for the jth examinee. As can be seen, the class membership probabilities are assumed to be affected by the covariates, which are typically modeled by specifying a logistic regression model for $\pi_{k|\underset{\sim}{\mathrm{z_j}}}$. That is,

$$\pi_{k|\underset{\sim}{\mathrm{z_j}}} = \frac{exp(\alpha_k + \sum_{p=1}^{P} \beta_{pk} z_{jp})}{\sum_{k'=1}^{K} exp(\alpha_{k'} + \sum_{p=1}^{P} \beta_{pk'} z_{jp})}, \tag{4.8}$$

where α_k and β_{pk} are the intercept and slope coefficients, respectively, for LC k. Based on the coefficients β_{pk}, the statistical significance of the covariate z_{jp} predicting the LC proportions can be examined. The MM-IRT-C model can be used to examine overall DIF, both observed and unobserved. Unlike IRT and MM-IRT approaches, which independently investigate observed and unobserved DIF, respectively, the MM-IRT-C model can be used to analyze both types of DIF simultaneously. There are several important reasons for doing so. First, we can test whether the occurrence of observed DIF may be attributable to more nuanced, unobserved DIF. As an example, one may obtain observed DIF on gender, but these DIF effects may be accounted for by some LCs with unobserved DIF on the item of interest. A second reason for using MM-IRT-C is that DIF on multiple observed characteristics may be accounted for by unobserved measurement groups. Third, because not all observed DIF can be accounted for by unobserved measurement groups, it is necessary to examine residual observed DIF beyond that of unobserved DIF. Conceptually, unobserved differences (i.e., latent class differences) may not fully demarcate how examinees differentially use a scale; such differences may be attributable to observed group membership (see Figure 1C, Paths 2 and 3). In this case, the conditional probability $P(y_{ji} \mid k, \theta_j)$ in equation 4.5 becomes $P(y_{ji} \mid k, \theta_j, \underset{\sim}{\mathrm{z_j}})$.



1 Mean-level differences on; latent trait between observed groups there needs to be at least one non-invariant item for a model with Path 1 estimated to be identified.

2 Nonuniform observed DIF (differences in item discriminations)

3 Uniform observed DIF (differences in item locations)

4 Mean-level differences on latent trait between unobserved groups; there needs to be at least one non-invariant item for a model with Path 4 estimated to be identified.

5 Nonuniform unobserved DIF (differences in item discriminations)

6 Uniform unobserved DIF (differences in item locations)

7 Covariate predicts latent class proportion

Figure 4.1: Graphical conceptualization of the MM-IRT-C model.

Source: Tay et al. (2011).

4.2 Equating using the IRT-C model

In this section we are going to explain how to conduct IRT true-score equating and IRT observed-score equating using the IRT-C model described in the previous section. We are going to use the 3PLM because it is the most general model.

4.2.1 IRT true-score equating with IRT-C

If we consider the pseudo-guessing parameter c_i^*, the IRT-C model presented in Equation 4.4 becomes (Tay et al., 2015):

$$P(y_{ji} \mid \theta_j, z_j) = c_i^* + \frac{1 - c_i^*}{1 + exp(-[a_i\theta_j + b_i + c_i z_j + d_i z_j \theta_j])}. \tag{4.9}$$

For this model, the number-correct true score on Form X which was described in subsection 2.3 is defined as:

$$\tau_X = \sum_{i:X} p_{ji}(\theta_j, z_j; a_i, b_i, c_i^*, c_i, d_i). \tag{4.10}$$

Likewise, the number-correct true score on Form Y is defined as

$$\tau_Y = \sum_{i:Y} p_{ji}(\theta_j, z_j; a_i, b_i, c_i^*, c_i, d_i). \tag{4.11}$$

The Form Y true score equivalent of a given true score on Form X is then defined as

$$irt_Y(\tau_X) = \tau_Y(\tau_X^{-1}). \tag{4.12}$$

4.2.2 IRT observed-score equating with IRT-C

Define $f_t(x \mid \theta_j, z_j)$ as the distribution of number-correct scores over the first t items for examinees of ability θ_j and covariate z_j (it was defined previously, but without covariates, in section 2.3). Define $f_1(x = 0 \mid \theta_j, z_j) = (1 - p_{j1})$ as the probability of earning a score of 0 on the first item and $f_1(x = 1 \mid \theta_j, z_j) = p_{j1}$ as the probability of earning a score of 1 on the first item. For $t > 1$, the recursion formula 2.108 for the IRT-C model 4.9 becomes:

$$\begin{aligned} f_t(x \mid \theta_j, z_j) &= f_{t-1}(x \mid \theta_j, z_j)(1 - p_{jt}), \quad x = 0, \\ &= f_{t-1}(x \mid \theta_j, z_j)(1 - p_{jt}) + f_{t-1}(x - 1 \mid \theta_j, z_j)p_{jt}, \quad 0 < x < t, \\ &= f_{t-1}(x - 1 \mid \theta_j, z_j)p_{jt}, \quad x = t. \end{aligned} \tag{4.13}$$

The recursion formula gives the observed score distribution for examinees of a given ability and covariate. To find the observed score distribution for examinees of various abilities and with different values of covariate, the observed score distribution for examinees at each ability and value of covariate is found and then these are accumulated. When the ability distribution is continuous, then

$$f(x) = \sum_{z_j} \int f(x \mid \theta, z_j)\psi(\theta)\, d\theta p(z_j), \tag{4.14}$$

where $\psi(\theta)$ is the distribution of θ and $p(z_j)$ is the distribution of z_j. To conduct observed score equating, observed score distributions are found for Form X and for Form Y. Conventional equipercentile methods are then used to find score equivalents.

4.3 A Simulation Study

The objective of the simulation is to show that the proposed methods for IRT equating with covariates can increase the accuracy of an equating by reducing the standard error of the estimators. In this section we present the set up of the simulations together with the obtained results which we obtained using the proposed method. It is important to underline that we used the bootstrap method because it is a very common approach in equating (Kolen and Brennan, 2014).

4.3.1 Simulation Set Up

To conduct the simulation, parametric bootstrap (Efron and Tibshirani, 1993) was used, especially the following steps:

1. It is assumed that, for the form X, probabilities p_{ji} are generated from the IRT-C model with fixed values of the parameters;
2. It is assumed that, for the form Y, probabilities p_{ji} are generated from the IRT-C model with fixed values of the parameters;
3. Randomly select N_X x n_X probabilities from the model from step 1 (where N_X is the number of the examinees and n_X the number of items on Form X);
4. Randomly select N_Y x n_Y probabilities from the model from step 2 (where N_Y is the number of the examinees and n_Y the number of items on Form Y);
5. Conduct IRT true score equating and IRT observed score equating using the parametric bootstrap samples drawn in steps 3 and 4;
6. Repeat steps 3 through 5 a large number of times R. The estimated standard error for the IRT true score equating is:

$$\hat{se}_{boot}[I\hat{R}T_Y(\tau_X)] = \sqrt{\frac{\sum_r [I\hat{R}T_{Yr}(\tau_X) - I\hat{R}T_{Y.}(\tau_X)]^2}{R-1}}, \tag{4.15}$$

where $I\hat{R}T_{Yr}(\tau_X)$ indicates the r bootstrap estimate and

$$I\hat{R}T_{Y.}(\tau_X) = \sum_r \frac{I\hat{R}T_{Yr}(\tau_X)}{R}. \tag{4.16}$$

The estimated standard error for the IRT observed score equating is:

$$\hat{se}_{boot}[\hat{e}_Y(x_i)] = \sqrt{\frac{\sum_r [\hat{e}_{Yr}(x_i) - \hat{e}_{Y.}(x_i)]^2}{R-1}}, \tag{4.17}$$

where $\hat{e}_{Yr}(x_i)$ indicates the r bootstrap estimate and

$$\hat{e}_{Y.}(x_i) = \sum_r \frac{\hat{e}_{Yr}(x_i)}{R}. \tag{4.18}$$

The idea is to compare standard errors obtained with this procedure with those obtained by the same procedure in which, however, at the first step the parameters are estimated using the classical IRT models (for example the 3PLM). If the standard errors obtained using the procedure presented above are lower than the standard errors obtained using the 3PLM, then we can conclude that the use of covariates actually increases the accuracy of the IRT equating.

4.3.2 Covariates

In order to simulate the relationships between the external covariates and the latent trait that have high fidelity to test data, we use the National Evaluation Institute for the School System (Invalsi) theta scores as simulated ability estimates. In particular, we use theta scores from the population of 593,407 examinees taking the 2013 Invalsi eight grade mathematics test. In this study, we used Invalsi theta scores as a proxy for the IRT latent trait scores because IRT scores have high correlations with the observed scores (Bartolucci et al., 2015).

Invalsi conducts annual surveys at different school grades. Invalsi develops tests to assess examinees' Italian language and mathematics competencies, and administers them to the whole population of primary school examinees (second and fifth grade), lower secondary school examinees (sixth and eighth grade), and upper secondary school examinees (tenth grade). We consider data from the administration of the Invalsi test at eighth grade. The choice of this grade assessment data is motivated by the fact that examinees receive a test score based on their performance, which contributes to the definition of the final examinee score at the end of lower secondary school within a state certification exam with legal validity. The choice of the mathematics test is motivated by the fact that Italian examinees have often shown serious gaps in this subject, even in international surveys (OECD, 2012). The 2013 Invalsi eigth grade math test consists of 45 dichotomous items divided into four areas: Numbers, Space and Figures, Data and Forecasts and Relations and Functions.

The covariates used were the gender and the nationality of the examinees. The choice of these covariates is motivated by the fact that girls and foreign citizens appear to be disadvantaged in

Table 4.1: Test Score Means for the Different Subpopulations in the Data from the 2013 Invalsi Eight Grade Math Test.

	Gender	
Nationality	MALE	FEMALE
ITA	24.74	23.20
FEFG	23.58	18.32
FESG	22.56	20.67

the mathematics test (see, for example, Tonello (2011)). An example of the relationship between the two covariates gender and nationality and the total score (between 0 and 50) is shown in Table 4.1, where ITA represents Italian examinees and $FEFG$ and $FESG$ represent, respectively, foreign examinees of first generation (examinees born abroad whose parents were both born abroad) and second generation (examinees born in Italy whose parents were both born abroad). The mean test score was higher for males than for females and also higher for Italian examinees than for foreign examinees with significant differences (INVALSI, 2013).

To predict the Invalsi theta scores (θ), an additive regression model was fit to the data with predictor variables of gender and nationality (where nationality was dummy coded):

$$\begin{aligned}\theta = 0.224 - 0.182(\text{FEMALE}) - 0.579(\text{FEFG})\\ - 0.274(\text{FESG}) + e,\end{aligned} \tag{4.19}$$

where e represents an error term. All the coefficients were significantly different from zero.

4.3.2.1 Design and Ability distributions

In this study we use the NEC design as the single group design is rarely used in practice because of order effects (Kolen and Brennan, 2014, p.14). We cannot use the NEAT design because there is no anchor test in the Invalsi test and there are not examinees with scores from two test administrations that we can use to create an hypothetic anchor. Finally, we do not use the EG design because it is reasonable to think that the examinees that sustained different test forms are not equivalent. For the NEC design, we need two samples: one sample with scores on test form X and one sample with scores on test form Y. For both forms, to simulate an ability distribution that conforms to the estimated additive model, random normal distribution were simulated for each of the six cells (gender x nationality) shown in Table 4.2 following the set up of Tay et al. (2015). All the variances of the normal distributions were set to 1, but the mean of the random normal distributions was dependent on the beta coefficients, as follows

Table 4.2: Percentages of examinees by gender and nationality.

%	ITA	FEFG	FESG	TOTAL
MALE	45.3	3.0	2.1	50.4
FEMALE	44.7	2.8	2.1	49.6
TOTAL	90.0	5.8	4.2	100.0

$$\begin{aligned}\theta^{*} mean = 0.224 - 0.182(\text{FEMALE}) - 0.579(\text{FEFG}) \\ - 0.274(\text{FESG}).\end{aligned} \tag{4.20}$$

The number of simulated examinees in each cell was based on the proportions in Table 4.2 multiplied by the total number of simulated examinees. We use the same model to simulate the ability distributions for the forms X and Y with the exception that the parameters of the IRT-C models will be different for the two forms.

4.3.2.2 Item parameters

In this study our objective is to show that the use of covariates can increase the accuracy of an equating. For this reason, in the simulation we use the IRT-C model, ignoring the latent classes, which are not our priority. Following Tay et al. (2015), for a traditional 3PLM, item discriminations a_i were sampled from a truncated normal distribution (mean=1.2, sd=0.3) with lowest and highest possible values set to 0.5 and 1.7 respectively; b_i were sampled from a uniform $(-2, 2)$ distribution and c_i^* were sampled from a logit-normal $(-1.1, 0.5)$ distribution.

4.3.2.3 DIF across multiple covariates

To simplify our simulation, we seek to only examine the presence of uniform DIF (Tay et al., 2015). We simulate DIF on the first fifteen (this number is not random: we explain this choice later) items. For moderate DIF, as it can be seen through the design matrix 4.21, we specified uniform DIF of the magnitude 0.40 on items $1, 2, 3, 4, 5$ and 12 and $14, 15$ biased against females. Items $6, 7, 8, 9, 10, 11$ and $14, 15$ are biased in favor of foreigns of first generation and, finally, items 11 and $13, 14, 15$ are biased against foreigns of second generation (in this last case, we specified uniform DIF of the magnitude 0.20). For large DIF, as it can be seen through the design matrix 4.22, we specified uniform DIF of the magnitude 0.60 in place of 0.40, and 0.30 in place of 0.20 (in specifying the magnitudes of 0.40 and 0.20 for moderate DIF and of 0.60 and 0.30 for large DIF we followed Tay et al. (2015)). Finally for low DIF, as it can be seen through the design matrix 4.23, we specified uniform DIF of the magnitude 0.10 in place of 0.40, and 0.05 in place of 0.20.

$$\begin{bmatrix} -0.40 & 0 & 0 \\ -0.40 & 0 & 0 \\ -0.40 & 0 & 0 \\ -0.40 & 0 & 0 \\ -0.40 & 0 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0.40 & -0.20 \\ -0.40 & 0 & 0 \\ 0 & 0 & -0.20 \\ -0.40 & 0.40 & -0.20 \\ -0.40 & 0.40 & -0.20 \end{bmatrix} \begin{bmatrix} \text{FEMALE} \\ \text{FEFG} \\ \text{FESG} \end{bmatrix}. \tag{4.21}$$

$$\begin{bmatrix} -0.60 & 0 & 0 \\ -0.60 & 0 & 0 \\ -0.60 & 0 & 0 \\ -0.60 & 0 & 0 \\ -0.60 & 0 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0.60 & 0 \\ 0 & 0.60 & -0.30 \\ -0.60 & 0 & 0 \\ 0 & 0 & -0.30 \\ -0.60 & 0.60 & -0.30 \\ -0.60 & 0.60 & -0.30 \end{bmatrix} \begin{bmatrix} \text{FEMALE} \\ \text{FEFG} \\ \text{FESG} \end{bmatrix}. \tag{4.22}$$

$$\begin{bmatrix} -0.10 & 0 & 0 \\ -0.10 & 0 & 0 \\ -0.10 & 0 & 0 \\ -0.10 & 0 & 0 \\ -0.10 & 0 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0.10 & 0 \\ 0 & 0.10 & -0.05 \\ -0.10 & 0 & 0 \\ 0 & 0 & -0.05 \\ -0.10 & 0.10 & -0.05 \\ -0.10 & 0.10 & -0.05 \end{bmatrix} \begin{bmatrix} \text{FEMALE} \\ \text{FEFG} \\ \text{FESG} \end{bmatrix}. \tag{4.23}$$

We underline that these matrixes are used for the purposes of simulation and do not reflect how or if the test is biased against different groups of examinees. The matrixes are the same for the two forms in this simulation only for simplicity, but we could have different matrixes or have no DIF on all of the items.

4.3.2.4 Test length and Sample Size

Test length should be at least $30 - 40$ items when equating educational tests with tables of specifications that reflect multiple areas of content (Kolen and Brennan, 2014, p.285). Following this rule, we simulated test lengths of 30 and 45 items where the first 15 items have DIF of the form specified in the DIF matrix but the remaining items do not have DIF. As such, our simulations focus on the boundary conditions with tests that have a large proportion of DIF items (50 percent) and a moderately large proportion (about 33 percent) of DIF items, respectively (Tay et al., 2015). The boundary conditions were introduced by Tay et al. (2015) to show that the IRT-C procedure for detecting DIF across multiple covariates is powerful because it works well even with a large proportion of DIF items; in this context it is useful to focus on these conditions to assess the impact of a different proportion of DIF items on equating. The need to focus on these boundary conditions respecting the rules for choosing an appropriate test length explains the choice to simulate exactly fifteen DIF items. Regarding the sample size, for the 3PLM $1,500$ examinees per form are usually required (Kolen and Brennan, 2014, p.304), so, presuming that we need a sample size of 500 to estimate each parameter, in our case a reasonable sample size could be 2,000 per form (in effect we have to estimate the three parameters of the 3PL IRT model and the DIF coefficient c_i). To evaluate how the method works with small samples, we also use the sample sizes of 200 and 600 per form.

4.3.3 Results

4.3.3.1 IRT observed-score equating

Using a moderate magnitude for the DIF, the IRT observed-score equating with IRT-C method already shows excellent results with small samples ($N1 = N2 = 200$), as can be seen in Figures A.1 and A.3. When we use a moderately large proportion of DIF items (33%) and the number of items is $n1 = n2 = 45$, we can observe in Figure A.1 that all the models with covariates have a lower SEE than the model without covariates over a large range of test scores. The SEE of the models with covariates is quite similar to the SEE of the model without covariates at the lower and at the upper end of the scale (Figure A.2), where, in general, a larger SEE is expected considering the fact that there are few examinees with the lowest results and few with the highest results. The model with only the covariate gender shows the worst performance among the models with covariates; however, it has a lower SEE than the model without covariates over a quite large range of test scores. When we use a large proportion of DIF items (50%) and the number of items is $n1 = n2 = 30$, we can observe in Figure A.3 that the models with covariates (except for the model with only the covariate gender) have a lower SEE than the model without covariates over a large range of test scores, even at the upper end of the scale. We can observe in Figure A.4 that the SEE of the models with covariates is quite similar to the SEE of the model without covariates at the lower end of the scale. The model with only the covariate gender shows a worse performance compared to the previous case. Figures A.5 and A.7 show the results obtained with medium samples ($N1 = N2 = 600$). When the number of items is $n1 = n2 = 45$, we can observe in Figure A.5 that all the models with covariates have a lower SEE than the model without covariates over a large range of test scores. As can be seen in Figure A.6, the model with only the covariate gender has an higher SEE than the model without covariates at the lower end of the scale, but however it has a lower SEE than the model without covariates over a quite large range of test scores. When the number of items is $n1 = n2 = 30$, we can observe in Figure A.7 that the models with covariates (except for the model with only the covariate gender) have a lower SEE than the model without covariates over the whole range of test scores, even at the lower and at the upper end of the scale (Figure A.8). Figures A.9 and A.11 show the results which we obtained with large samples ($N1 = N2 = 2000$). When the number of items is $n1 = n2 = 45$, we can observe in Figure A.9 that all the models with covariates have a lower SEE than the model without covariates over the whole range of test score, even at the lower and at the upper end of the scale (Figure A.10). When the number of items is $n1 = n2 = 30$, we can observe in Figure A.11 that the models with covariates (except for the model with only the covariate gender) have a lower SEE than the model without covariates over the whole range of test scores, even at the lower and at the upper end of the scale. The model with only the covariate gender shows a worse performance compared to the previous case and it has an higher SEE than the model without covariates at the lower and at the upper end of the scale (Figure A.12).

Using a large magnitude for the DIF, we obtain results which are very similar to the ones shown in Figures A.1, A.3, A.5, A.7, A.9 and A.11, as we can see in Figures A.13-A.18.

Using a low magnitude for the DIF and small samples ($N1 = N2 = 200$), we obtained results which are very similar to the ones shown in Figures A.1 and A.3, as we can see in Figures A.21 and A.22. When the number of items is $n1 = n2 = 45$, we can observe in Figure A.21 that the models with covariates (except for the three models with only one covariate) have a lower SEE than the model without covariates over the whole range of test scores, even at the lower and at the upper end of the scale (Figure A.22). When the number of items is $n1 = n2 = 30$, we can observe in Figure A.23 that the models with covariates (except for the two models with only the covariates

GEN and FEFG, respectively) have a lower SEE than the model without covariates over the whole range of test scores, even at the lower and at the upper end of the scale (Figure A.24). When the number of items is $n1 = n2 = 45$, we can observe in Figure A.25 that the models with covariates (except for the three models with only one covariate) have a lower SEE than the model without covariates over the whole range of test scores, even at the lower and at the upper end of the scale (Figure A.26). When the number of items is $n1 = n2 = 30$, we can observe in Figure A.27 that the models with covariates (except for the two models with only the covariates GEN and FEFG, respectively) have a lower SEE than the model without covariates over a quite large range of test scores. At the lower end of the scale several models with covariates have an higher SEE than the model without covariates (Figure A.28).

Summing up, IRT observed-score equating showed excellent results: in fact, in general, all the models with covariates had a lower SEE than the model without covariates. The method already worked well with small samples ($N1 = N2 = 200$) and the moderate or large entity of the DIF did not have a relevant impact on the results. A low entity of the DIF shows instead a worse performances of the models with only one covariate. When the number of items is $n1 = n2 = 30$, we observed a better performance of the models with covariates than the model without covariates even at the lower and at the upper end of the scale.

4.3.3.2 IRT true-score equating

The IRT true-score equating with IRT-C method did not show good results. Figures 4.2 shows the results obtained with small samples ($N1 = N2 = 200$), when the number of items is $n1 = n2 = 30$. Using a large DIF and a number of items of $n1 = n2 = 45$, we obtained results which are very similar to the ones shown in Figure 4.2, for this reason we omitted that graph. Increasing the sample size, we cannot improve our results. For instance, Figures 4.3 shows the results which we obtained for the models without covariates (red line) and with only the covariate FESG (blue line) using big samples ($N1 = N2 = 2000$), when the number of items is $n1 = n2 = 45$.

Summing up, we cannot recommend to use IRT true-score equating with covariates, because the models with covariates did not have a lower SEE than the model without covariates. Increasing the sample size, the entity of the DIF or the number of the items, we cannot improve our results. A possible reason of these negative results could be that true scores are not avalaible in practice and no justification exists for applying the true score relationship to observed scores.

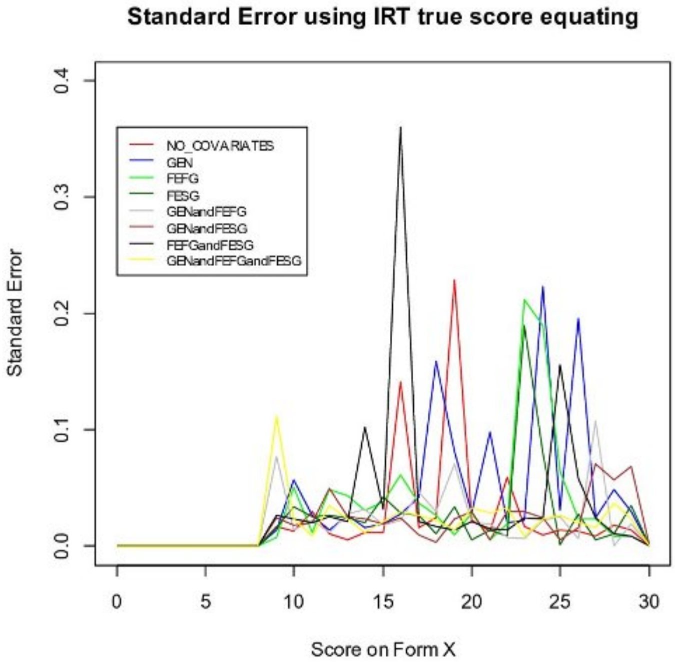


Figure 4.2: SSE using IRT true-score equating when $n1 = n2 = 30$, $N1 = N2 = 200$ and DIF is moderate.

4.4 An Empirical Study

4.4.1 Introduction

To illustrate the use of the techniques presented earlier with real data, we equated two administrations, 2012 and 2013, of the Invalsi eighth grade math test. We have already described the test administrated in 2013 (see section 4.3). The 2012 Invalsi eighth grade mathematics test consists of 46 dichotomous items divided into four areas: Numbers, Space and Figures, Data and Forecasts and Relations and Functions. The Invalsi test is given once a year, with a new test form every time: because for the government and for the schools is essential to understand if and how the knowledge of the subject (mathematics in our example) changes for a specific grade (the eighth grade in our example), there is therefore a need to equate test scores from different test forms. So, in this context we are in an horizontal equating situation (see subsection 1.1.3). Even if we do not focus on a vertical equating situation here, we want to underline that for the institutions is also essential to understand if and how the knowledge of the subject changes across grades, so in the future Invalsi will have to face this issue).

In this testing situation, none of the requirements for standard equating are fulfilled. There is no anchor test (all items are released after the administration of the test), and there may be differences in ability between the examinee groups, which makes the assumption of random samples from the

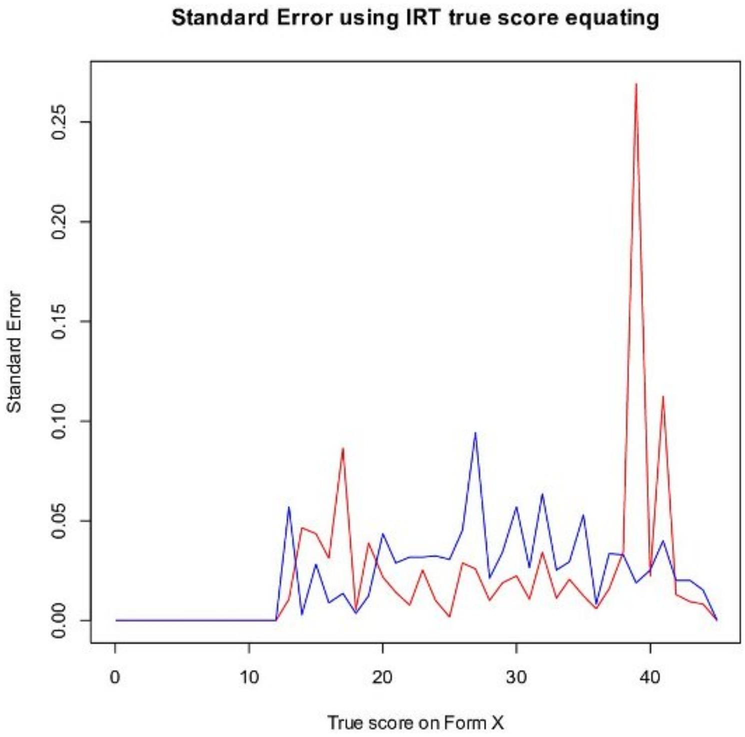


Figure 4.3: SSE of the models without covariates (red line) and with only the covariate FESG (blue line) using IRT true-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is moderate.

same population implausible. To adjust for these differences, we can use covariates thus we will compare the results of the EG design with the results of the NEC design. The covariates used were gender (GEN) (which was coded 0 for males and 1 for females), foreign examinee of first generation (FEFG) and foreign examinee of second generation (FESG). The number of examinees taking the 2012 administration was $N = 587,412$, and the number of examinees taking the 2013 administration was $N = 593,407$. Due to these large sample sizes, we choose to use Trento's data (with sample sizes $N = 5,029$ in 2012 and $N = 4,998$ in 2013). The choice of Trento is motivated by the fact that this Italian province showed significantly different Invalsi math scores between males and females and between Italian examinees and foreign examinees in 2013 (INVALSI, 2013). It is very important to underline that Trento does not show atypical distributions of the covariates. The distributions are similar to the distributions of the covariates of the other Italian regions, as we can see in Tables 4.3 and 4.4.

4.4.2 Results

To estimate item parameters, DIF and examinees' ability parameters, we used the R packages difR (Magis et al., 2015) and ltm (Rizopoulos, 2006). We can see the estimates of the item parameters and of the DIF for the 2012 and 2013 Trento's data in Tables 4.5 - 4.8. For the 2012 Trento's data we can see in Table 4.7 that there is no DIF on the following items: 7,10,12 and 32 for the covariate GEN; 1,4,24 and 26 for the covariate FEFG and 16,24,43 and 45 for the covariate FESG. We also see in Table 4.7 that there is a very large DIF (> 0.60) on the following items: 3,5 and 46 (biased in favor of females) and 15,16,25,26,36 and 37 (biased against females) for the covariate GEN; 2,6,8,12,14,17,19,25 and 37 (biased in favor of foreigns of first generation) and 9,10,16,29,36,38,39,41 and 45 (biased against foreigns of first generation) for the covariate FEFG; 3,12 and 14 (biased in favor of foreigns of second generation) and 8,21,28 and 33 (biased against foreigns of second generation) for the covariate FESG. Excluding the items without DIF and with a very large DIF, all the other items have the following mean DIF (calculated on the absolute values): 0.40, 0.30 and 0.30 for the covariates GEN, FEFG and FESG, respectively. Compared to the thresholds used in the simulation study, we can say that the mean DIF across the three covariates (calculated excluding the extreme values) has medium-high magnitude. For the 2013 Trento's data we can see in Table 4.8 that there is no DIF on the following items: 4,5,8,21 and 43 for the covariate GEN and 1,20 and 31 for the covariate FEFG (there is DIF on all of the items for the covariate FESG). We also see in Table 4.8 that there is a very large DIF (> 0.60) on the following items: 15,16,25,27,28 and 36 (biased in favor of females) and 7,13,23,29 and 30 (biased against females) for the covariate GEN; 15,16,24,27,28,33,44 and 45 (biased in favor of foreigns of first generation) and 4,5,18,21,22,35,36 and 40 (biased against foreigns of first generation) for the covariate FEFG; 24,27,28,33 and 45 (biased in favor of foreigns of second generation) and 6,11,26,36,37,40 and 41 (biased against foreigns of second generation) for the covariate FESG. Excluding the items without DIF and with a very large DIF, all the other items have a mean DIF (calculated on the absolute values)of 0.30 for the three covariates. So also in this case, compared to the thresholds used in the simulation study, we can say that the mean DIF across the three covariates (calculated excluding the extreme values) has medium-high magnitude. For the sake of brevity, we omit the estimates of the examinees' ability parameters: we only say that the ability parameters' means for the 2012 and 2013 Trento's data are, respectively, $3e-05$ and 0.000665. We can see the results which we obtained applying the IRT observed-score equating with covariates on the Trento's data in Figures 4.4. Substantially, all the models with covariates have a lower SEE than the model without covariates over the whole range of test scores, even at the upper end of the scale. We can observe in Figure 4.5 that the models with covariates have quite similar SEE than the model without covariates at the lower end of the scale. In Figure 4.6 we can see that the difference between the equated scores and the X scores are larger using the EG design.

Table 4.3: Distribution of the gender in the Invalsi data for the different Italian regions for the year 2013, percentages.

	Male	Female	Missing
Valle d'Aosta	49.2	50.7	0.1
Piemonte	50.1	49.6	0.3
Liguria	50.6	49.1	0.3
Lombardia	50.1	49.8	0.1
Veneto	50.0	50.0	0.0
Friuli-VG	50.2	49.8	0.0
Emilia Romagna	50.0	49.7	0.3
Toscana	50.5	49.0	0.5
Umbria	51.6	48.4	0.0
Marche	50.7	49.1	0.2
Lazio	49.9	49.6	0.5
Abruzzo	50.1	49.9	0.0
Molise	49.2	50.7	0.1
Campania	50.6	49.1	0.3
Basilicata	50.6	49.1	0.3
Calabria	50.5	49.3	0.2
Sicilia	49.6	50.2	0.2
Sardegna	49.4	49.1	1.5
Bolzano italiana	52.5	47.3	0.2
Bolzano ladina	49.8	50.2	0.0
TRENTO	49.5	50.4	0.1

Table 4.4: Distribution of the foreign examinees of first generation (FEFG) and of second generation (FESG) in the Invalsi data for the different Italian regions for the year 2013, percentages.

	Other	FEFG	FESG	Missing
Valle d'Aosta	88.9	6.3	4.6	0.2
Piemonte	85.9	8.5	5.0	0.6
Liguria	86.7	8.7	4.0	0.6
Lombardia	84.6	8.8	6.3	0.3
Veneto	86.3	8.3	5.2	0.2
Friuli-VG	86.1	8.1	5.6	0.2
Emilia Romagna	83.7	9.4	6.3	0.6
Toscana	80.0	8.4	11	0.6
Umbria	84.2	10.1	5.6	0.1
Marche	86.9	7.35	5.5	0.25
Lazio	88.6	6.6	4.1	0.7
Abruzzo	90.8	5.1	3.8	0.3
Molise	94.3	3.5	2.1	0.1
Campania	97.0	1.4	1.0	0.6
Basilicata	96.6	1.5	1.4	0.5
Calabria	95.5	2.7	1.2	0.6
Sicilia	96.1	1.7	1.6	0.6
Sardegna	95.6	1.4	1.2	1.8
Bolzano italiana	74.5	16.8	8.1	0.6
Bolzano ladina	97.7	1.9	0.4	0.0
TRENTO	86.5	8.2	5.2	0.1

Table 4.5: Estimates of the item parameters for the 2012 Trento's data.

	a	b	c
item 1	0.739	-5.603	4.209e-02
item 2	0.531	-0.318	3.490e-01
item 3	0.544	-3.245	4.367e-01
item 4	1.309	0.339	1.818e-01
item 5	0.486	-2.894	1.626e-01
item 6	1.958	1.535	2.631e-01
item 7	0.969	-0.052	3.862e-01
item 8	1.434	1.124	3.883e-01
item 9	0.719	-1.614	3.749e-04
item 10	0.649	-2.393	1.731e-03
item 11	1.450	0.625	2.747e-01
item 12	1.914	2.331	1.717e-01
item 13	2.446	0.465	2.574e-01
item 14	1.561	1.847	2.990e-01
item 15	1.224	0.917	3.561e-02
item 16	1.557	1.059	1.158e-01
item 17	2.001	1.771	3.611e-01
item 18	1.127	0.422	1.922e-01
item 19	0.969	2.837	1.972e-01
item 20	1.548	1.702	6.048e-02
item 21	0.892	-0.295	4.205e-05
item 22	0.920	0.368	6.151e-05
item 23	1.163	1.603	1.586e-01
item 24	1.515	1.643	3.349e-02
item 25	2.527	1.525	2.166e-01
item 26	2.103	0.683	1.665e-01
item 27	1.857	0.510	2.238e-01
item 28	1.805	0.714	7.063e-02
item 29	1.689	0.126	1.914e-01
item 30	1.555	0.135	2.289e-01
item 31	1.925	0.595	2.507e-01
item 32	1.039	1.307	1.033e-01
item 33	0.902	-1.020	3.831e-01
item 34	1.664	1.216	3.363e-01
item 35	0.986	-0.577	3.017e-01
item 36	1.050	0.392	9.194e-05
item 37	1.178	0.864	2.092e-01
item 38	2.461	-1.076	2.889e-05
item 39	1.638	-0.606	5.979e-07
item 40	1.254	-0.241	1.102e-05
item 41	1.855	-0.770	2.249e-07
item 42	0.931	0.1432	3.884e-01
item 43	0.941	-0.201	5.113e-02
item 44	1.090	-0.478	2.814e-01
item 45	1.321	0.026	2.839e-02
item 46	0.759	-1.397	6.170e-03

Table 4.6: Estimates of the item parameters for the 2013 Trento's data.

	a	b	c
item 1	0.784	-1.277	6.959e-02
item 2	0.958	-0.517	1.157e-01
item 3	1.833	0.277	2.866e-01
item 4	2.224	0.333	8.539e-02
item 5	2.204	0.332	7.56e-03
item 6	0.860	-1.067	1.046e-01
item 7	0.865	0.284	1.532e-04
item 8	0.591	-0.151	9.799e-06
item 9	0.670	-0.454	1.322e-02
item 10	1.528	0.192	1.336e-01
item 11	1.269	-0.146	1.611e-01
item 12	1.014	0.707	5.129e-04
item 13	1.305	0.947	1.389e-01
item 14	1.001	-2.086	2.273e-05
item 15	0.950	1.618	3.162e-01
item 16	0.245	-5.092	6.806e-03
item 17	1.409	0.156	1.621e-01
item 18	1.452	0.187	4.057e-04
item 19	1.529	-0.037	2.681e-01
item 20	1.128	0.553	3.017e-04
item 21	2.020	-0.285	1.716e-01
item 22	1.776	0.179	2.315e-06
item 23	1.390	1.888	1.198e-04
item 24	1.196	1.146	1.165e-01
item 25	0.891	-0.603	3.669e-09
item 26	1.265	0.0466	5.681e-01
item 27	-1.800	-2.478	3.780e-01
item 28	-0.051	-0.651	1.341e-02
item 29	1.161	-0.091	2.738e-03
item 30	1.486	0.392	4.509e-02
item 31	0.800	0.472	1.454e-03
item 32	1.293	0.140	1.359e-01
item 33	0.293	-2.782	2.195e-04
item 34	0.949	0.619	2.260e-04
item 35	1.380	-0.183	1.431e-01
item 36	0.826	0.006	2.350e-04
item 37	1.971	0.246	1.962e-01
item 38	1.467	0.3761	4.360e-02
item 39	0.747	-0.191	1.924e-04
item 40	1.482	0.211	7.677e-07
item 41	1.321	0.532	1.820e-01
item 42	1.269	0.373	4.968e-02
item 43	1.940	0.6998	6.467e-01
item 44	1.431	1.203	4.698e-01
item 45	1.798	1.576	3.083e-01

Table 4.7: Estimates of the DIF of the covariates GEN, FEFG and FESG for the 2012 Trento's data.

	GEN	FEFG	FESG
item 1	0.2044	0.0362	0.4127
item 2	0.2261	0.7666	-0.1208
item 3	1.0843	0.5417	0.6681
item 4	-0.5748	-0.0209	-0.5692
item 5	1.1191	0.3128	0.1473
item 6	0.1006	0.7734	-0.0985
item 7	0.0174	-0.3722	-0.3830
item 8	-0.3053	0.7640	-0.7370
item 9	0.2825	-1.1155	-0.1476
item 10	-0.0287	-1.0180	0.3793
item 11	-0.1227	0.3553	0.1410
item 12	-0.0432	1.4500	0.8124
item 13	-0.5121	-0.1998	0.2080
item 14	0.2383	1.1336	0.7877
item 15	-0.7388	-0.1650	0.3010
item 16	-1.0664	-0.7454	-0.0207
item 17	-0.0501	0.8751	0.4625
item 18	-0.0606	-0.2284	0.1687
item 19	0.4481	1.3649	0.4712
item 20	-0.5772	0.4063	-0.3784
item 21	0.5327	-0.4672	-0.6947
item 22	0.6296	-0.2240	-0.5887
item 23	0.5746	-0.0966	0.0639
item 24	-0.3746	-0.0367	0.0371
item 25	-0.8728	0.9978	0.6077
item 26	-0.9623	-0.0322	-0.0754
item 27	-0.6138	-0.3283	-0.4510
item 28	-0.4813	-0.2113	-0.8869
item 29	-0.1929	-0.6904	-0.3196
item 30	-0.0693	0.1076	-0.1765
item 31	0.0517	0.1645	0.2707
item 32	-0.0031	0.1687	0.3039
item 33	0.3116	-0.2460	-0.7078
item 34	0.4065	0.5626	0.3833
item 35	0.4065	-0.1422	0.1532
item 36	-0.6665	-0.7842	0.2547
item 37	-0.9539	1.0878	0.3741
item 38	0.4453	-1.4140	-0.1422
item 39	0.4488	-0.9078	-0.3894
item 40	0.4705	-0.4094	-0.0733
item 41	0.5584	-0.8472	-0.1260
item 42	0.2171	0.1542	-0.2477
item 43	0.4026	-0.3163	0.0306
item 44	0.1974	0.1020	0.4808
item 45	-0.4197	-0.8749	0.0122
item 46	0.6806	-0.2317	-0.5988

Table 4.8: Estimates of the DIF of the covariates GEN, FEFG and FESG for the 2013 Trento's data.

	GEN	FEFG	FESG
item 1	-0.4818	0.0143	0.2047
item 2	-0.5351	-0.3445	-0.0823
item 3	-0.0921	-0.0808	-0.2590
item 4	-0.0078	-0.9525	-0.2049
item 5	0.0049	-1.4455	-0.4138
item 6	0.0736	-0.2794	-0.6547
item 7	-0.7271	-0.2221	-0.4308
item 8	-0.0303	-0.2883	0.5729
item 9	0.1457	0.1955	0.6058
item 10	0.1325	-0.5196	-0.5997
item 11	0.2726	-0.5006	-0.7734
item 12	0.1549	-0.1079	0.3692
item 13	-0.6951	0.6004	0.3718
item 14	0.3471	-0.5222	-0.2522
item 15	0.9275	1.2716	0.3652
item 16	0.7833	0.7457	0.4481
item 17	0.0959	-0.2042	0.1053
item 18	0.0994	-0.8481	-0.3457
item 19	-0.4214	-0.2037	0.0855
item 20	-0.3267	-0.0106	-0.6110
item 21	-0.0362	-0.8986	0.0529
item 22	-0.5313	-0.7104	-0.1224
item 23	-0.9553	-0.4709	0.3838
item 24	0.1692	0.8138	0.7828
item 25	0.9048	-0.2933	-0.0877
item 26	-0.3593	-0.1774	-0.7280
item 27	0.7607	2.0798	1.5660
item 28	0.9602	1.7916	1.2906
item 29	-1.1703	0.2230	-0.2700
item 30	-1.9813	0.3071	-0.1732
item 31	-0.1488	-0.0070	0.5882
item 32	0.1685	-0.2721	-0.5974
item 33	0.1058	0.7917	0.6968
item 34	-0.0801	0.4491	0.2409
item 35	0.6455	-0.7863	0.5038
item 36	0.9945	-0.7781	-0.7412
item 37	-0.5626	-0.2841	-0.8429
item 38	0.4872	-0.4926	-0.6239
item 39	0.4261	-0.6178	-0.1445
item 40	0.3835	-1.3348	-0.9240
item 41	0.4150	0.4305	-0.7330
item 42	-0.3553	0.0940	-0.2999
item 43	0.0425	0.8956	0.3046
item 44	0.3325	1.2441	0.4254
item 45	-0.3353	1.7061	0.9513

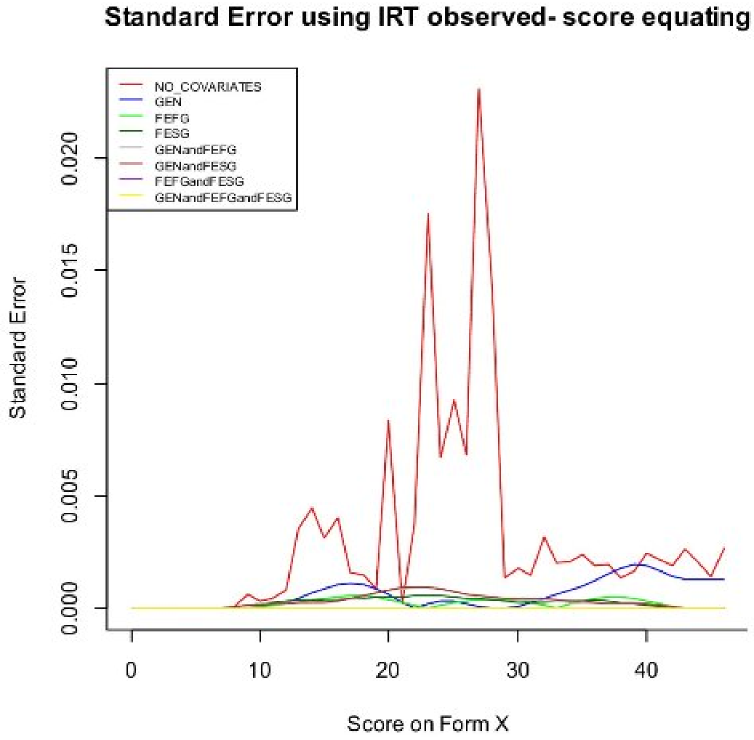


Figure 4.4: SSE using IRT observed-score equating for the Trento's data.

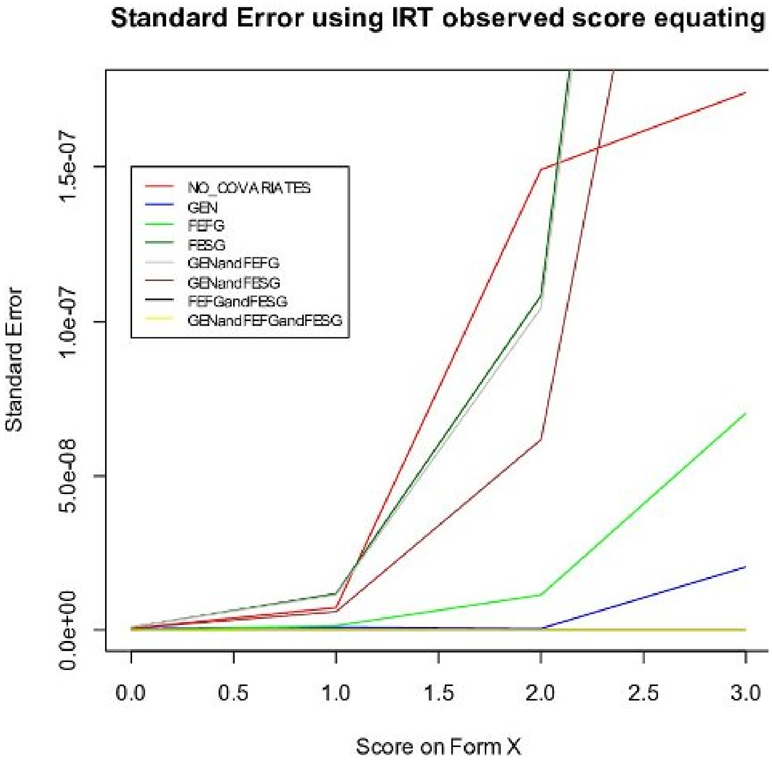


Figure 4.5: SSE at the lower end of the scale using IRT observed-score equating for the Trento's data.

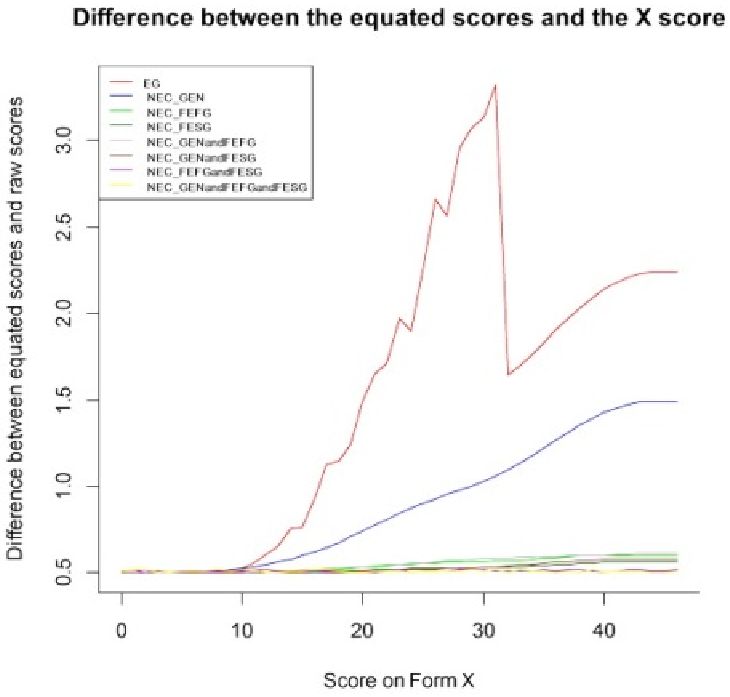


Figure 4.6: Difference between the equated scores and the X scores for the EG and the NEC design.

Chapter 5

Conclusions

This work has dealt with equating and, in particular, with equating using IRT.

In the first chapter a general overview of equating has been provided. Equating has been defined as the statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen and Brennan, 2014). Subsequently equating properties and designs have been considered in detail, because these concepts have been used as the principal basis for developing equating procedures. Then the standard error of equating has been defined as the standard deviation of score equivalents over replications of the equating procedure. Finally, the practical issues that are involved in conducting equating have been described.

In the second chapter the traditional equating methods using the equivalent groups and the NEAT designs have been described. Subsequently IRT equating and the kernel method of test equating have been illustrated in detail. Then two practical problems have been discussed: how to average results from more than one equating and how to conduct equating with small samples. Finally, it is explained how to evaluate the results of equating and some recent ideas in test equating have been described.

The third chapter has dealt with equating with covariates. It has been underlined that equating with covariates is motivated by an interest in the possibility of using information about the examinees' background to increase the precision of the estimators. Then it has been explained how to use covariates in observed score linear equating, kernel equating and in a Bayesian context.

The fourth chapter has proposed a new IRT true-score equating method and a new IRT observed-score equating method that can be used with the NEC design. The idea was to improve the accuracy of the equating by lowering the SE in the absence of an anchor test.

The proposed IRT true-score equating method, substantially, consists in finding the equivalent true scores on the equated forms using the IRT-C model (Tay et al., 2015). The proposed IRT observed-score equating method let us to obtain the observed score distribution for examinees of a given ability and covariate using an updated version of the recursion formula (Lord and Wingersky, 1984). Both the methods have been discussed together with simulations, while a real application has been done only for the IRT observed-score equating method.

Both the simulated data and the real test empirical data studies have been done using Invalsi data and the same covariates: the gender and the nationality of the examinees. To conduct the simulation, the following choices have been done:

- According to Kolen and Brennan (2014) the parametric bootstrap (Efron and Tibshirani,

1993) has been used;

- The Invalsi theta scores have been used as simulated ability estimates;
- The NEC and the EG designs have been used, because there is no anchor test in the Invalsi test, so it is impossible to use a NEAT design;
- Following Tay et al. (2015), item discriminations were sampled from a truncated normal distribution, item difficulties were sampled from a uniform distribution and the pseudo-guessing parameters were sampled from a logit-normal distribution;
- To simplify our simulation, only the presence of uniform DIF has been examined and DIF has been simulated only on the first fifteen items;
- Test lengths of 30 and 45 items have been simulated (Kolen and Brennan, 2014, p.285).

The results from the simulation study support the IRT observed-score equating method. The SEs obtained with the proposed method in a NEC design are lower in all examined conditions compared to using the traditional method in an EG design. These results are in line with those obtained by Wiberg and Bränberg (2015), although they examined the possible use of a NEC design within the kernel equating framework.

The results from the simulation study show that the IRT true-score equating method doesn't work: the models with covariates don't have a lower SE than the models without covariates and increasing the sample size, the entity of the DIF or the number of the items, we cannot improve our results. Due to these results, we cannot recommend to use IRT true-score equating with covariates.

To illustrate the use of the IRT observed-score equating method with real data, two administrations, 2012 and 2013, of the Invalsi eighth grade math test heve been equated. In particular, Trento's data (with sample sizes $N = 5,029$ in 2012 and $N = 4,998$ in 2013) has been used. The choice of Trento is motivated by the fact that this Italian province showed significantly different Invalsi math scores between males and females and between Italian examinees and foreign examinees in 2013 (INVALSI, 2013).

The real test data study strengthens the results from the simulation study and also shows that the proposed method can be used in practice. It was evident that the SEs were lower with the proposed method using a NEC design than using the traditional IRT observed-score equating with an EG design with the Invalsi data. We compared the results with the EG design because this is typically what we have access to if we do not have an anchor test, and this was also noted in Wiberg and Bränberg (2015). Due to these excellent results, we can recommend to use IRT observed-score equating with covariates.

A limitation of this study is our focus on SE: future studies should also examine bias. Of course, in such a case one runs into the problem of how to define a true equating function (Wiberg and González, 2016). Also, the choice of covariates in this study was guided by availability, and in the future one could examine other covariates. Finally, we only used dichotomous items and we presumed the existence of only one latent dimension underlying the data, and it could be interesting to extend the used model to the case of polytomous items as well as to presume the existence of more latent dimensions.

Bibliography

Albano, A. D. (2015). A general linear method for equating with small samples. *Journal of Educational Measurement*, 52(1):55–69.

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. Also available as `http://www.jstatsoft.org/v55/i06/`.

Andersson, B. and von Davier, A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement*, 51(3):223–238.

Babcock, B., Albano, A., and Raymond, M. (2012). Nominal weights mean equating: a method for very small samples. *Educational and Psychological Measurement*, 72:608–628.

Barrientos, A., Jara, A., and Quintana, F. (2012). Fully nonparametric regression for bounded data using bernstein polynomials. Technical report, Department of Statistics, Pontificia Universidad Catolica de Chile.

Bartolucci, F., Bacci, S., and Gnaldi, M. (2015). *Statistical analysis of questionnaires: a unified approach based on R and Stata*. Chapman and Hall/CRC.

Bränberg, K. and Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, 48(4):419–440.

Camilli, G., Wang, M., and Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, 32(1):79–96.

Cook, L. L., Eignor, D. R., and Schmitt, A. P. (1990). Equating achievement tests using samples matched on ability. College board report 90-2, New York, NY: College Entrance Examination Board.

Darroch, J. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, (43):1470–1480.

DiBello, L. V., Roussos, L. A., and Stout, W. (2007). *Handbook of statistics*, volume 26, chapter Review of cognitively diagnostic assessment and a summary of psychometric models, pages 979–1030. Amsterdam, the Netherlands: Elsevier.

Divgi, D. R. (1987). A stable curvilinear alternative to linear equating. Report crc 571, Alexandria, VA: Center for Naval Analyses.

Efron, B. and Morris, C. (1977). Steins paradox in statistics. *Scientific American*, (236):119–127.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman Hall.

Eid, M. and Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, (16):20–30.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum.

Gebhardt, E. and Adams, R. (2007). The influence of equating methodology on reported trends in pisa. *Journal of Applied Measurement*, 8(3):305–322.

González, J., Barrientos, A., and Quintana, F. (2015). *Quantitative Psychology Research*, chapter A dependent bayesian nonparametric model for test equating, pages 213–226. 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.* DOI: 10.1177/0146621616629380.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

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

Häggström, J. and Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement*, 51:201–211.

Hanson, B. and Béguin, A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1):3–24.

Harris, D., editor (1993). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Holland, P. and Rubin, D., editors (1982a). *Test equating*, chapter Observed score test equating: a mathematical analysis of some ETS equating procedures, pages 9–49. New York: Academic Press.

Holland, P. and Rubin, D., editors (1982b). *Test equating*, chapter On the foundations of test equating, pages 9–49. New York, NY: Academic Press.

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

Holland, P. W., Dorans, N. J., and S., P. N. (2007). *Handbook of statistics*, volume 26, chapter Equating test scores, pages 169–203. Oxford, England: Elsevier.

Holland, P. W. and Strawderman, W. (1989). The symmetric average of equating functions. Unpublished manuscript.

Hulin, C. L., Drasgow, F., and Parsons, C. K. (1983). *Item response theory: application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

INVALSI (2013). Rilevazioni nazionali sugli apprendimenti 2012-13. Technical report, INVALSI Publishing. Also available as `http://www.invalsi.it/snvpn2013/rapporti/Rapporto_SNV_PN_2013_DEF_11_07_2013.pdf/`.

Karabatsos, G. and Walker, S. (2009). A Bayesian nonparametric approach to test equating. *Psychometrika*, (74):211–232.

Kendall, M. G. and Stuart, A. (1977). *The advanced theory of statistics*. New York, NY: Macmillan, 4th edition.

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

Kolen, M. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 18:1–11.

Kolen, M. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, (9):25–44.

Kolen, M. (1990). Does matching in equating work?A discussion. *Applied Measurement in Education*, 3(1):23–39.

Kolen, M. and Brennan, R. (2014). *Test equating, scaling, and linking: methods and practices*. New York, NY: Springer-Verlag, 3nd edition.

Kyoung Yim, M. and Sun, H. (2006). Test equating of the medical licensing examination in 2003 and 2004 based on the item response theory. *JEEHP*, 3:2. Also available as `http://www.jeehp.org/DOIx.php?id=10.3352/jeehp.2006.3.2`.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka 's rule-space approach. *Journal of Educational Measurement*, 41:205–237.

Levine, R. (1955). Equating the score scales of alternative forms administered to samples of different ability. ETS research bulletin rb-55-23, Princeton, NJ: ETS.

Liang, T. and von Davier, A. (2014). Cross-validation: an alternative bandwidth-selection method in kernel equating. *Applied Psychological Measurement*, 38(4):281–295.

Linn, R., editor (1989). *Educational measurement*, chapter Scaling, norming and equating, pages 221–262. New York, NY: American Council on Education and Macmillan, 3rd edition.

Liou, M., Cheng, P. E., and Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, 25(2):197–207.

Lissitz, R. and Huynh, H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Also available as `http://pareonline.net/getvn.asp?v=8&n=10`.

Livingston, S. and Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46:330–343.

Livingston, S. and Kim, S. (2010). An empirical comparison of methods for equating with randomly equivalent groups of 50 to 400 test takers. ETS Research Rept. RR-10-05, Princeton, NJ: ETS.

Livingston, S. and Lewis, C. (2009). Small-sample equating with prior information. ETS Research Rept. RR-09-25, Princeton, NJ: ETS.

Livingston, S. A., Dorans, N. J., and Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1):73–95.

Lord, F. (1965). A strong true score theory with applications. *Psychometrika*, (30):239–270.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, (7):165–174.

Lord, F. M. and Wingersky, M. S. (1984). Comparison of irt true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, (8):452–461.

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

MacEachern, S. N. (2000). Dependent dirichlet processes. Technical report, Department of Statistics, The Ohio State University.

Magis, D., Beland, S., and Raiche, G. (2015). *difR: collection of methods to detect dichotomous differential item functioning (DIF).* R package version 4.6.

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

OECD (2012). Pisa 2012 results-Italy-OECD. Technical report, OECD Publishing. Also available as `http://www.oecd.org/pisa/keyfindings/PISA-2012-results-italy-ITA.pdf/`.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review, Otaru University of Commerce*, 51(1):1–23.

Orbanz, P. and Teh, Y. W. (2010). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. New York: Springer.

Potthoff, R. F. (1966). *Multivariate analysis: proceedings of an international symposium held in Dayton, Ohio, June 14-19, 1965*, chapter Equating of grades or scores on the basis of a common battery of measurements, pages 541–559. New York, NY: Academic Press.

Puhan, G., Moses, T. P., Grant, M. C., and McHale, F. (2009). Small-sample equating using a singlegroup nearly equivalent test (SIGNET) design. *Journal of Educational Measurement*, (46):344–362.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, (53):495–502.

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Scott, D. (1992). *Multivariate density estimation.* New York, NY: John Wiley.

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

Swaminathan, H. and Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, (27):361–370.

Tay, L., Newman, D., and Vermunt, J. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement Equivalence. *Organizational Research Methods*, 1(14):147–176.

Tay, L., Newman, D., and Vermunt, J. (2015). Item response theory with covariates (IRT-C): assessing item recovery and differential item functioning for the three-parameter logistic model (in press). *Educational and Psychological Measurement.*

Thorndike, R. L., editor (1971). *Educational measurement*, chapter Scales, norms and equivalent scores, pages 508–600. Washington DC: American Council on Education, 2nd edition.

Tonello, M. (2011). Peer effects between non-native and native students: first evidence from Italy. IZA Summer School.

von Davier, A., editor (2011). *Statistical models for test equating, scaling, and linking.* New York: Springer.

von Davier, A. (2013). Observed-score equating: an overview. *Psychometrika*, 78:605–623.

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer-Verlag.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, (54):426–482.

Walker, S. and Muliere, P. (2003). A bivariate Dirichlet process. *Statistics and Probability Letters*, (64):1–7.

Wang, T. (2008). The continuized log-linear method: an alternative to the kernel method of continuization in test equating. *Applied Psychological Measurement*, (32):527–542.

Wiberg, M. and Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, pages 1–13.

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. and van der Linden, W. (2011). Local linear observed-score equating. *Journal of Educational Measurement*, 48(3):229–254.

Wright, N. K. and Dorans, N. J. (1993). Using the selection variable for matching or equating. Research Report RR-93-04, Princeton, NJ: ETS.

Xin, T. and Jiahui, Z. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement*, 39(1):44–61.

()

Appendix A

SSE using IRT observed-score equating

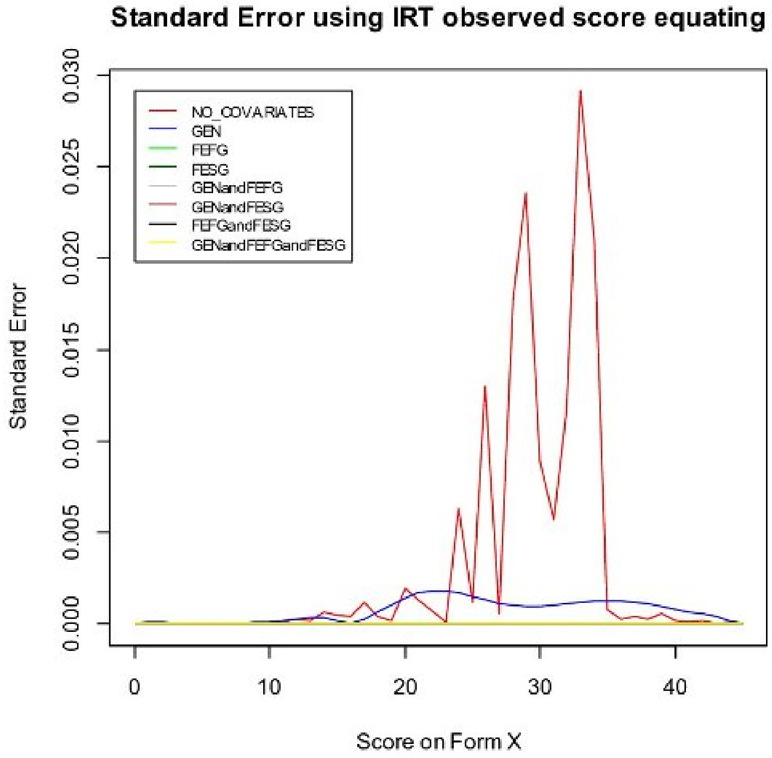


Figure A.1: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 200$ and DIF is moderate.

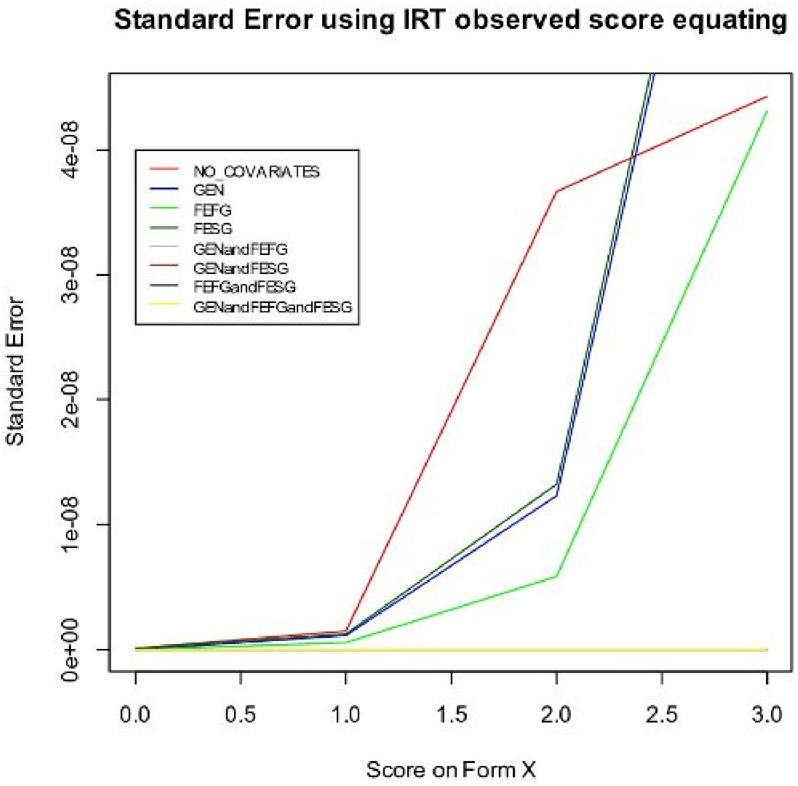


Figure A.2: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 200$ and DIF is moderate.

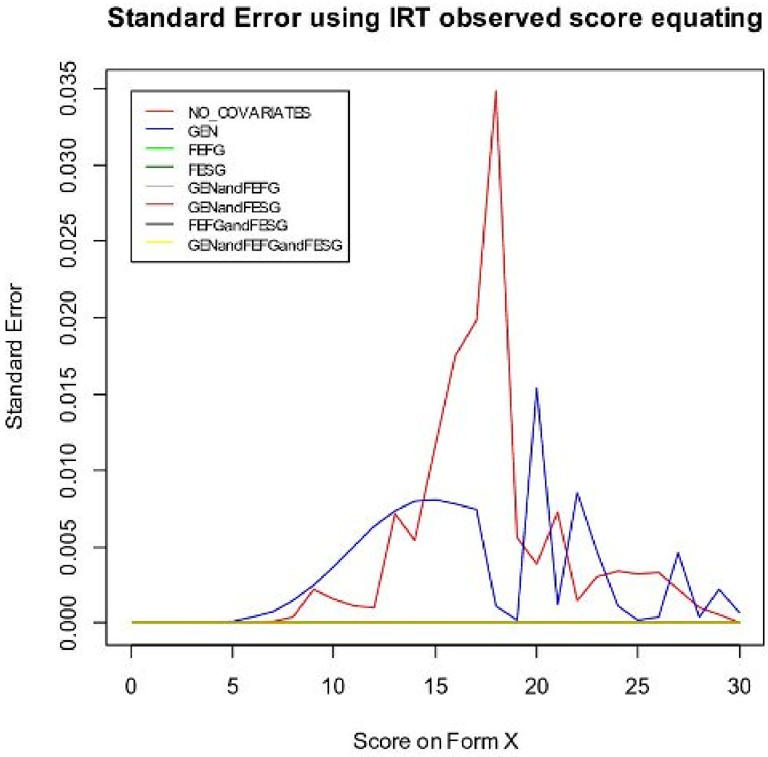


Figure A.3: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 200$ and DIF is moderate.

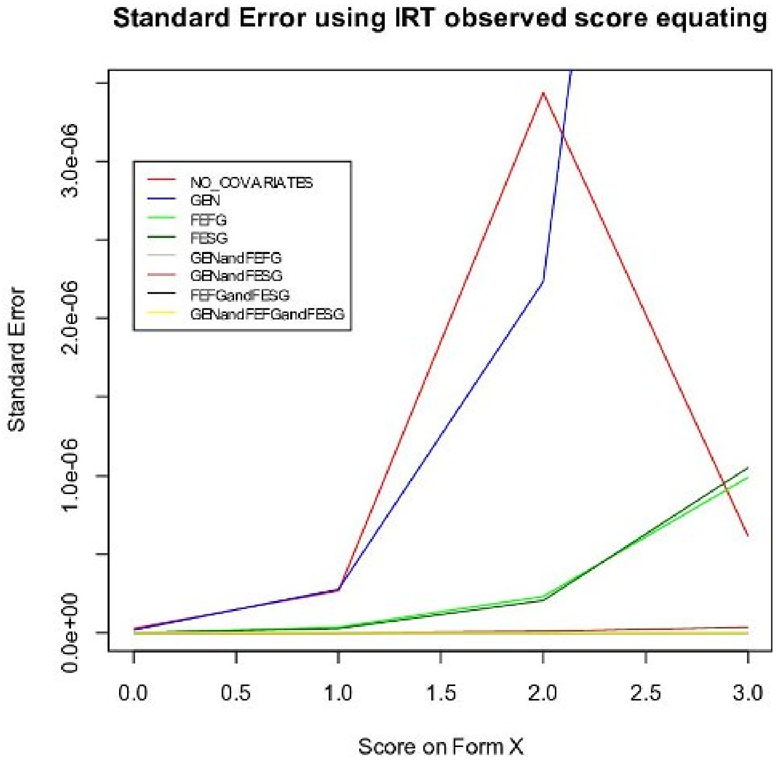


Figure A.4: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 200$ and DIF is moderate.

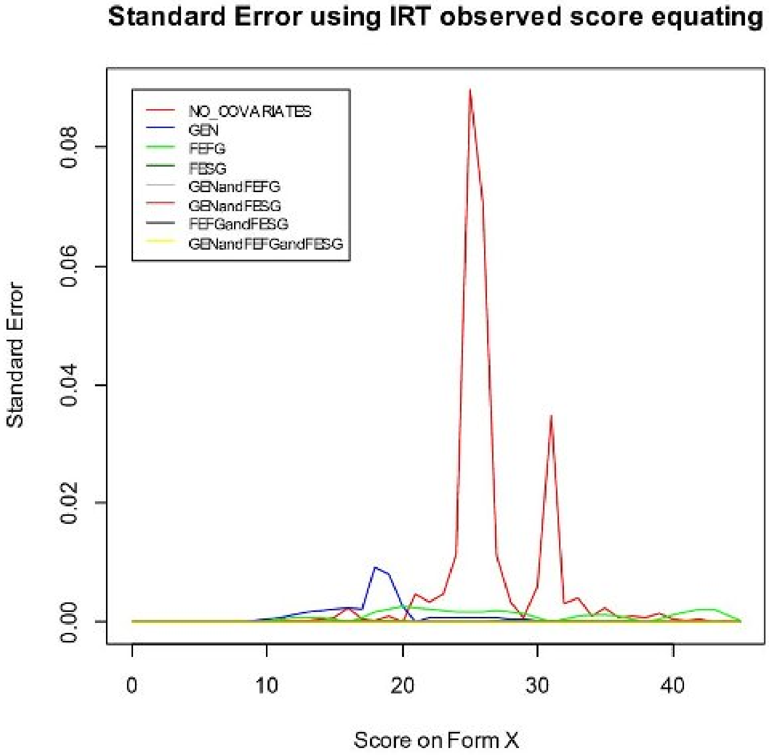


Figure A.5: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 600$ and DIF is moderate.

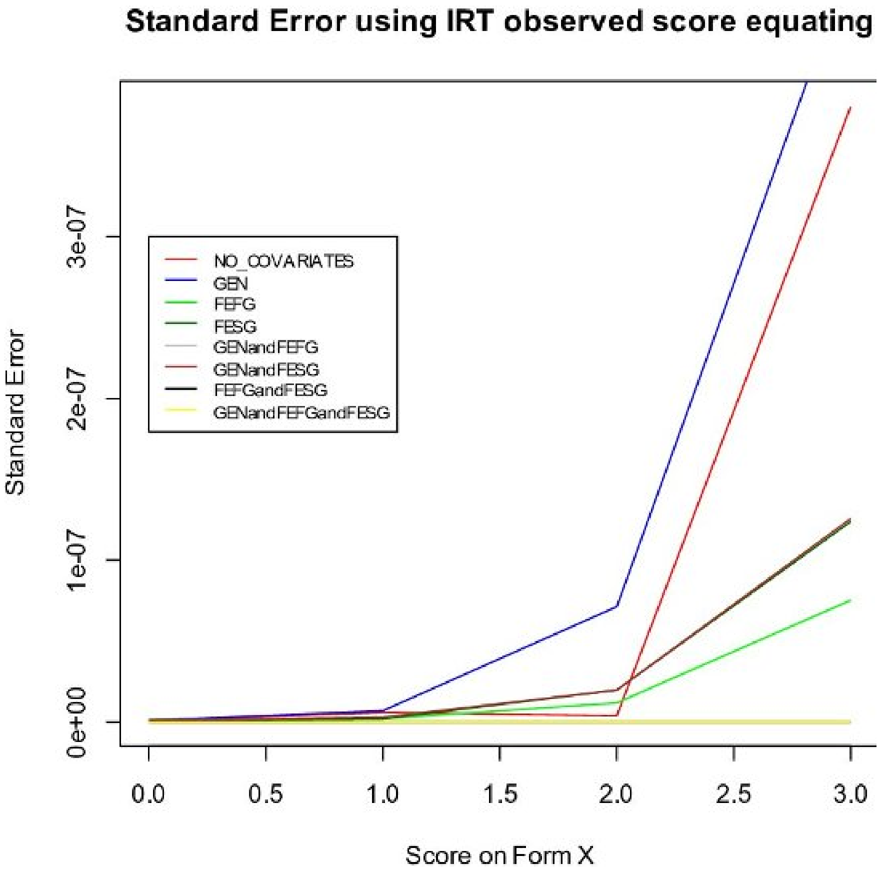


Figure A.6: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 600$ and DIF is moderate.

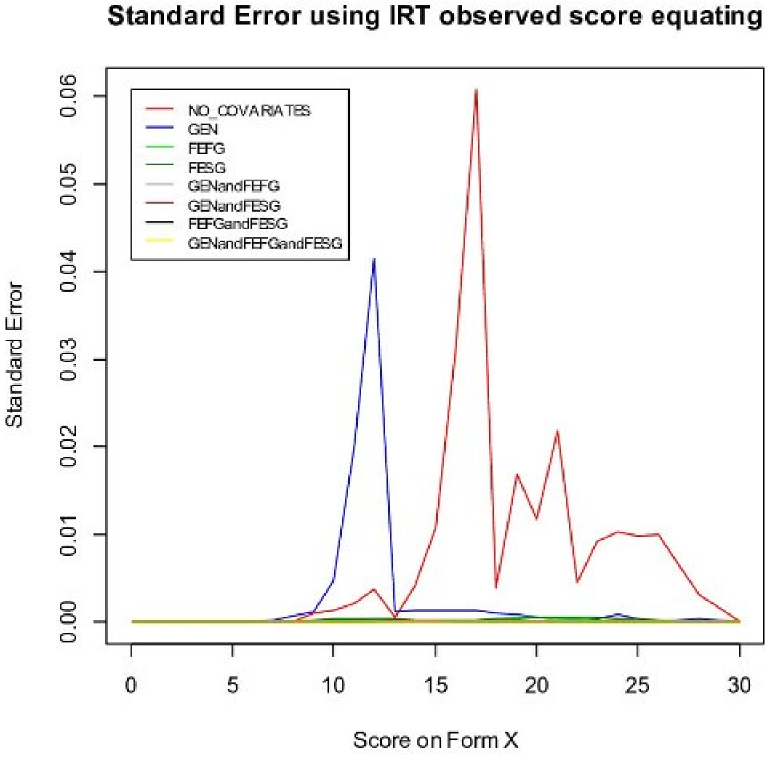


Figure A.7: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 600$ and DIF is moderate.

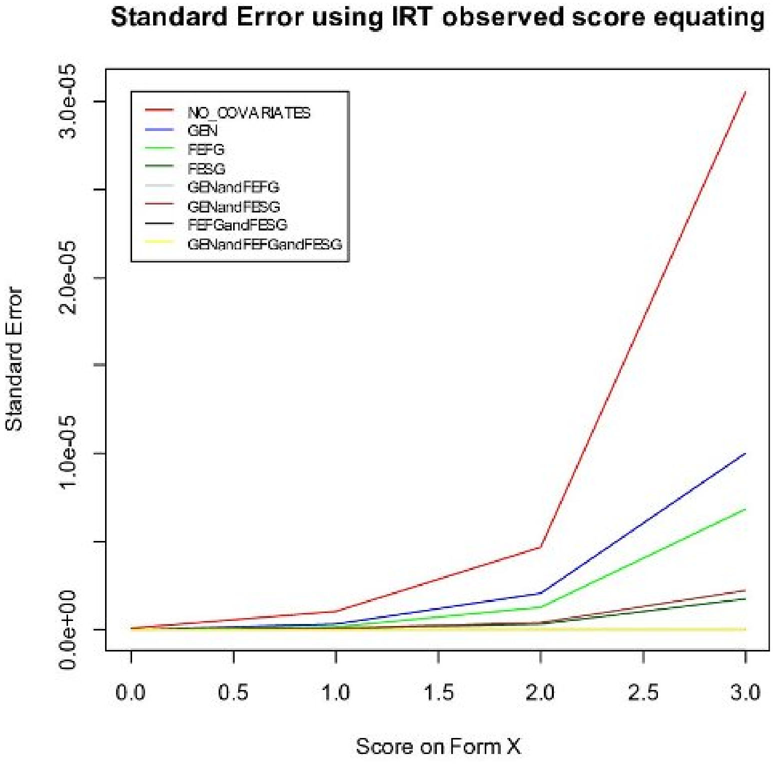


Figure A.8: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 600$ and DIF is moderate.

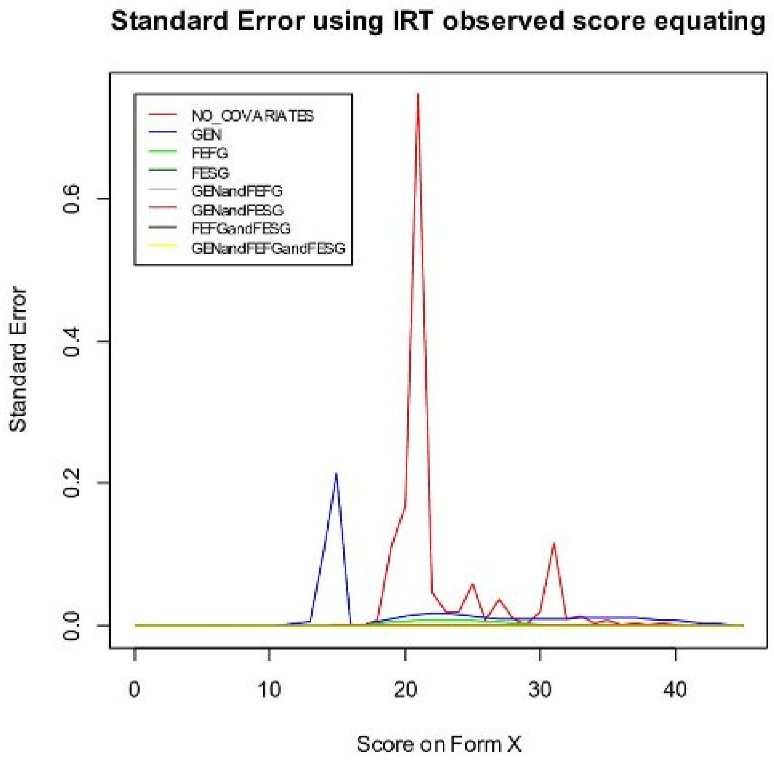


Figure A.9: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is moderate.

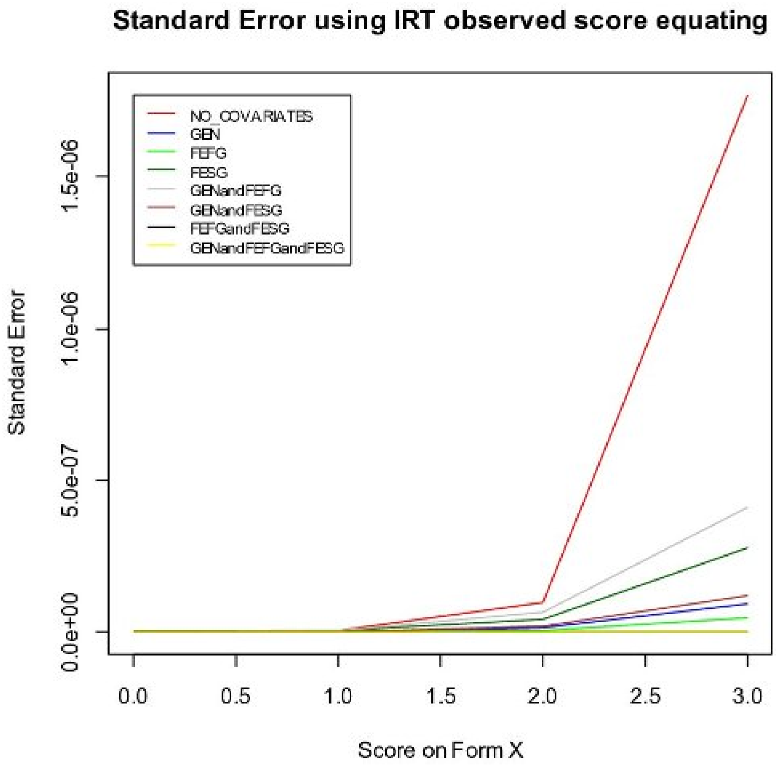


Figure A.10: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is moderate.

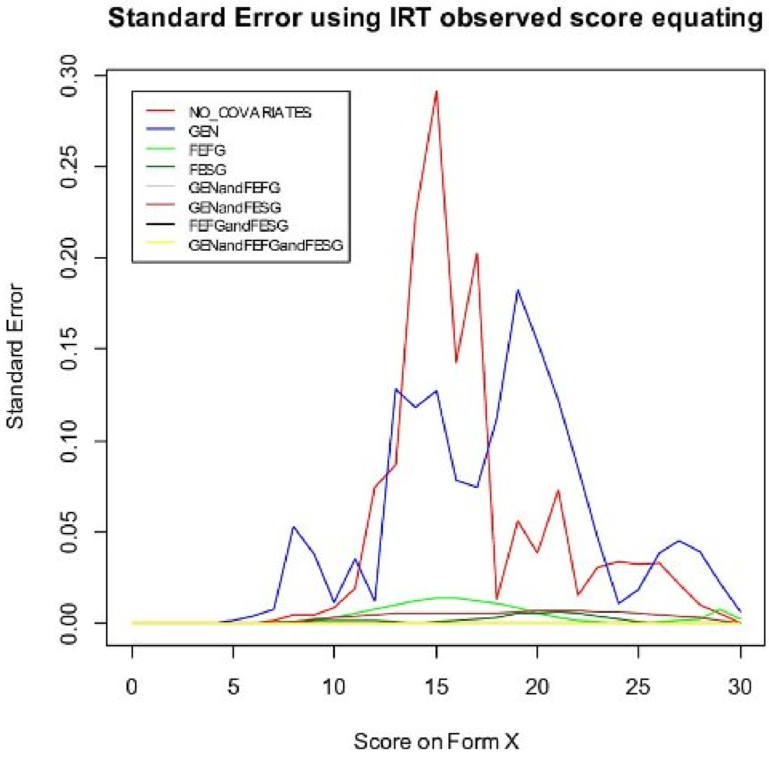


Figure A.11: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 2000$ and DIF is moderate.

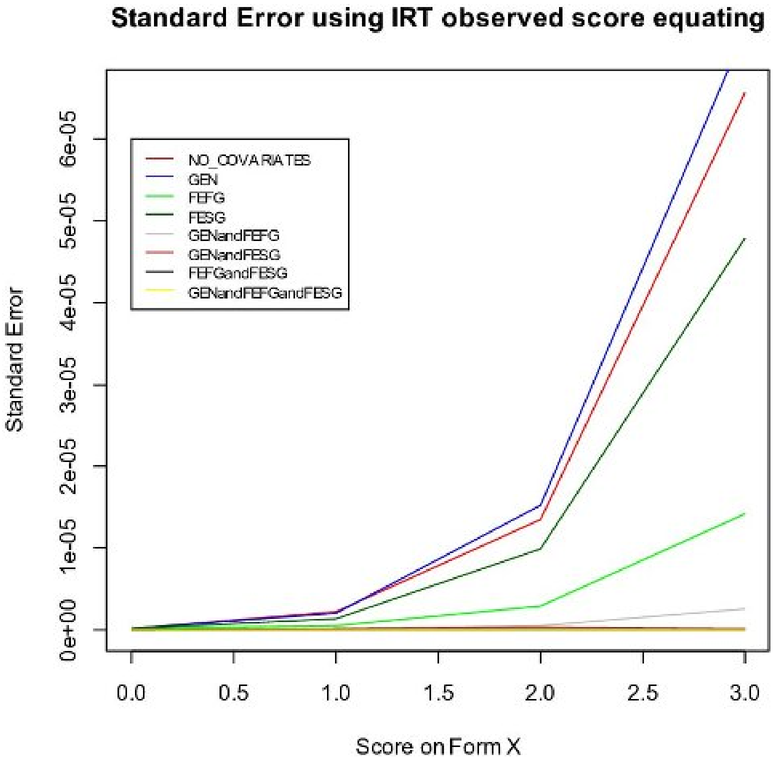


Figure A.12: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 2000$ and DIF is moderate.

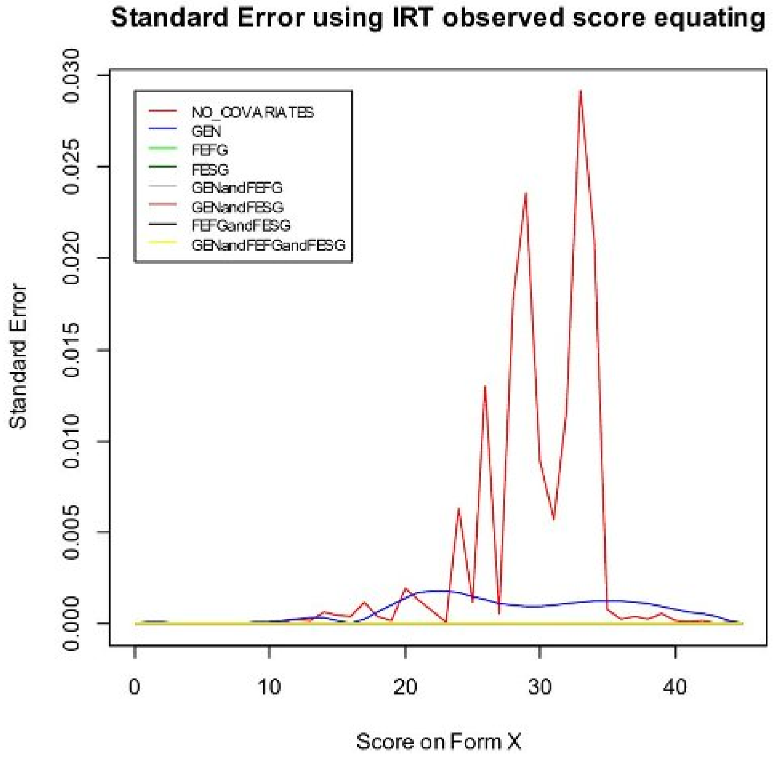


Figure A.13: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 200$ and DIF is high.

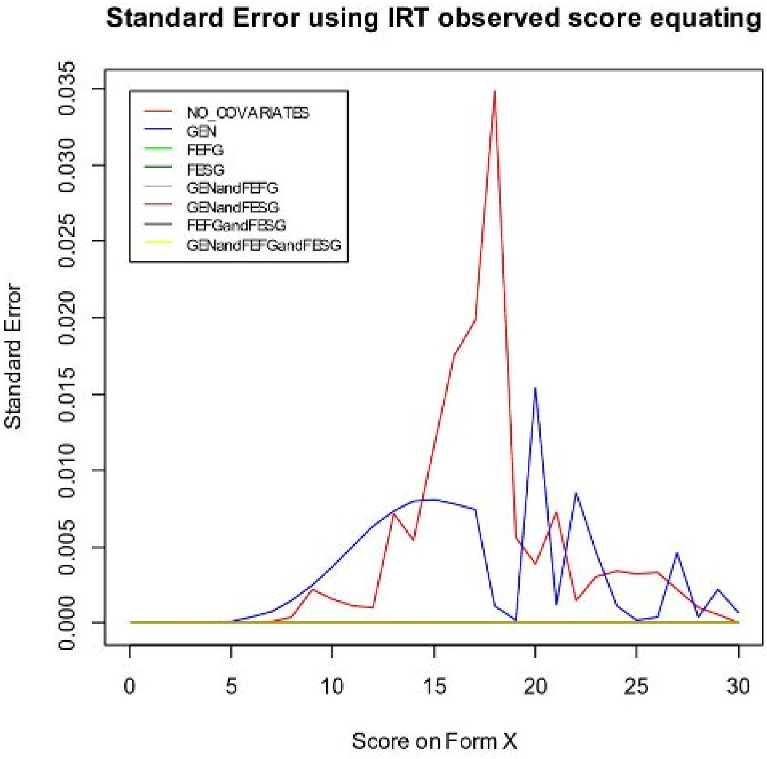


Figure A.14: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 200$ and DIF is high.

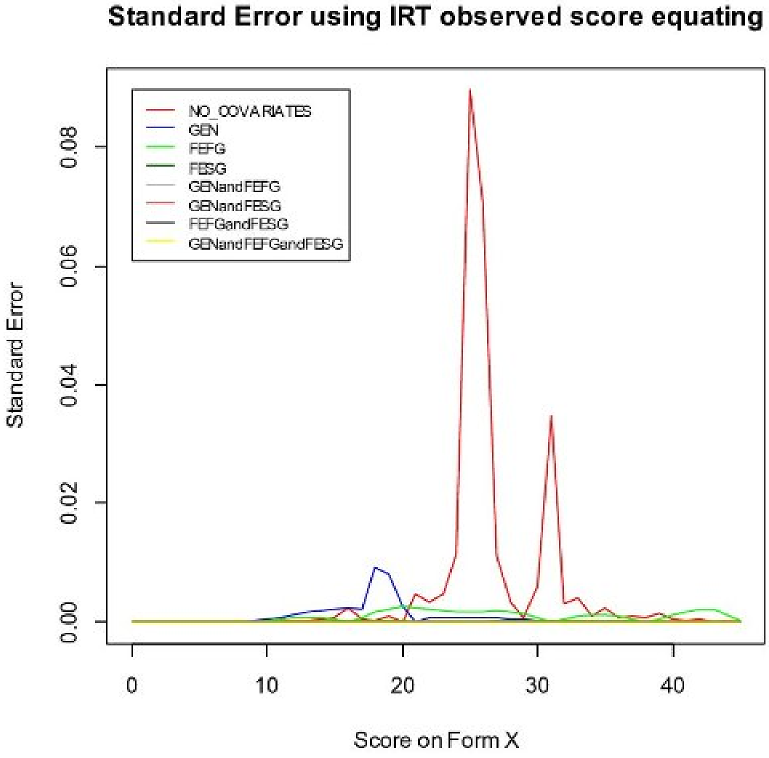


Figure A.15: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 600$ and DIF is high.

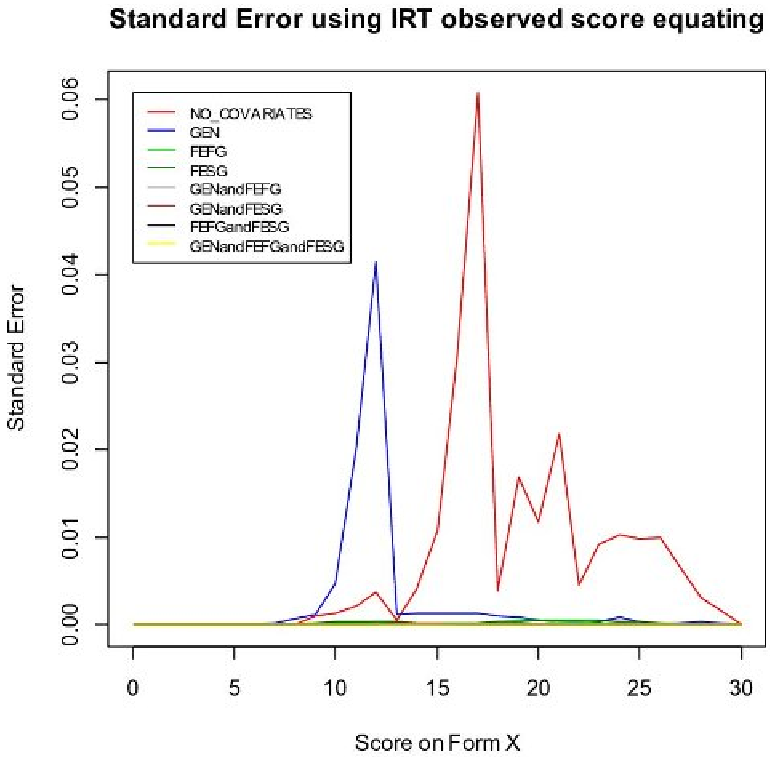


Figure A.16: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 600$ and DIF is high.

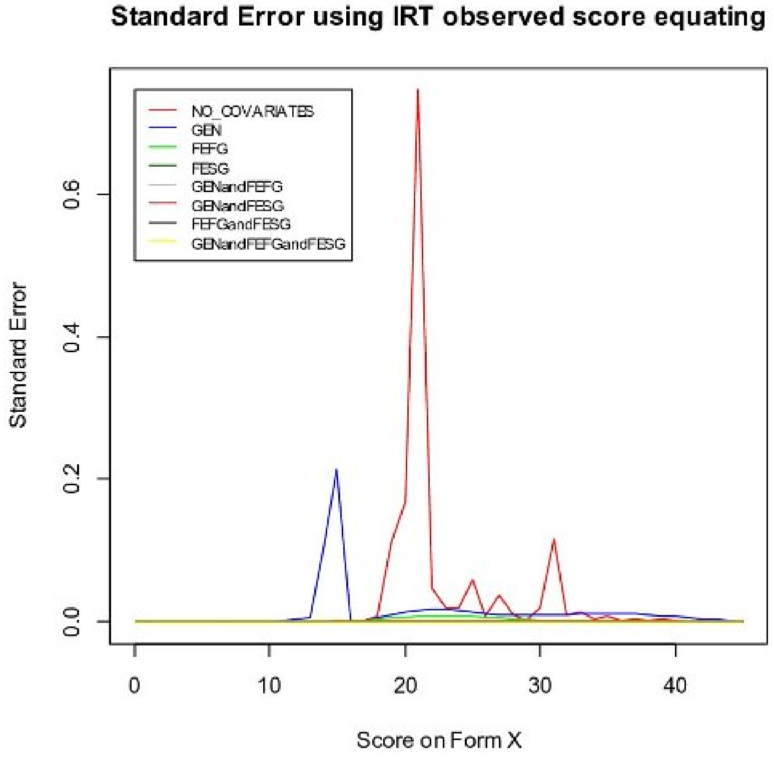


Figure A.17: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is high.

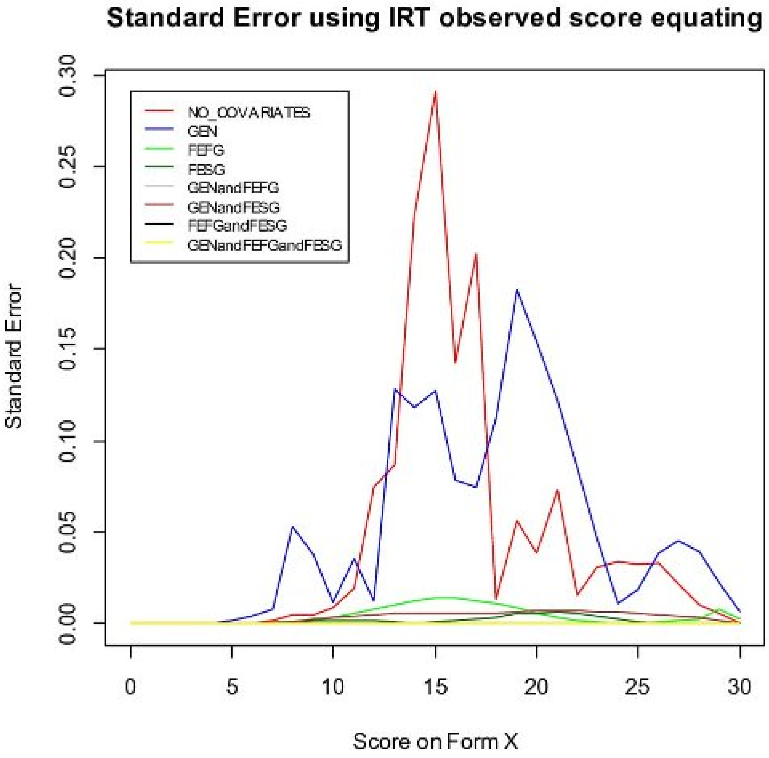


Figure A.18: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 2000$ and DIF is high.

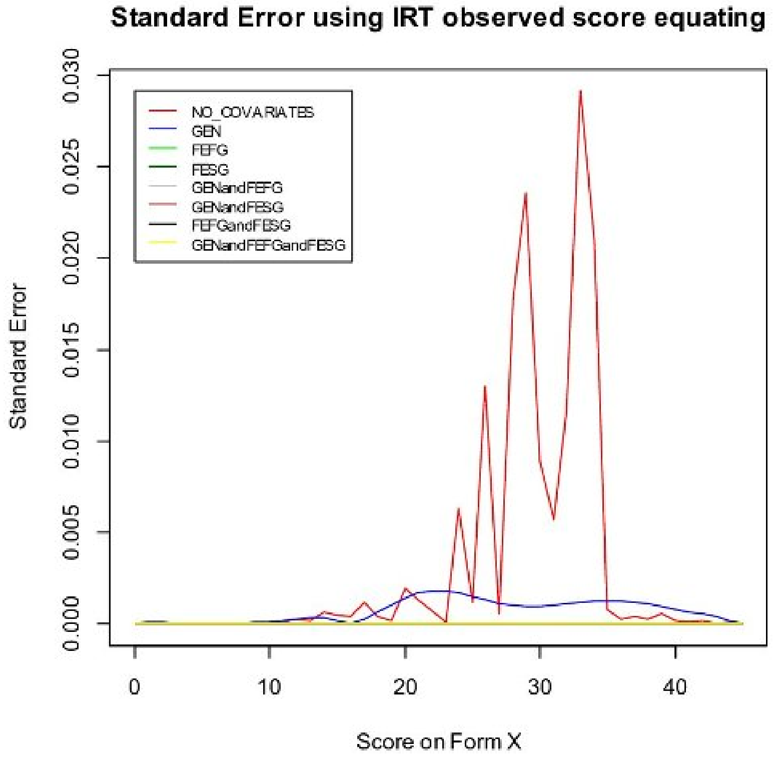


Figure A.19: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 200$ and DIF is low.

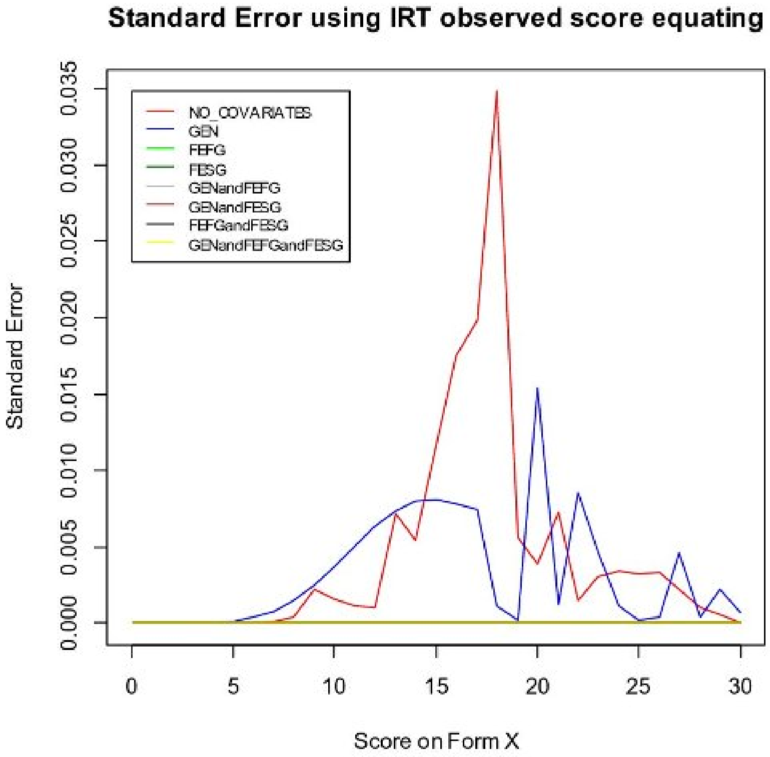


Figure A.20: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 200$ and DIF is low.

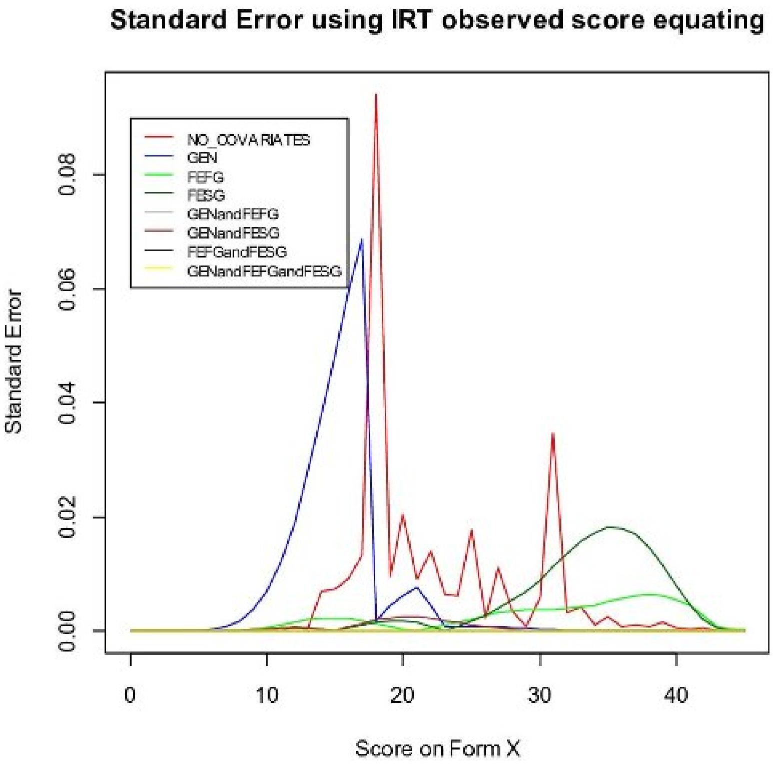


Figure A.21: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 600$ and DIF is low.

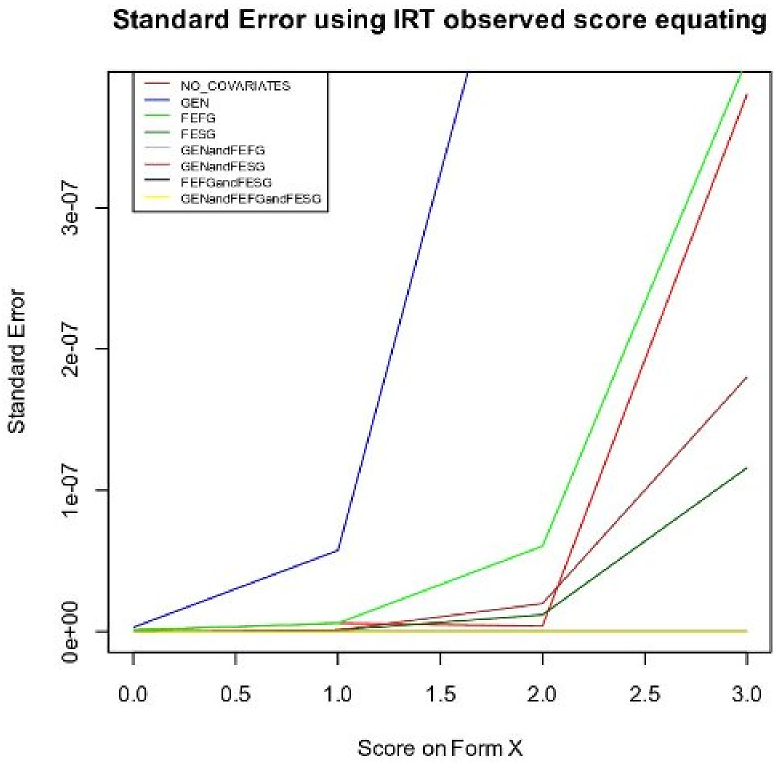


Figure A.22: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 600$ and DIF is low.

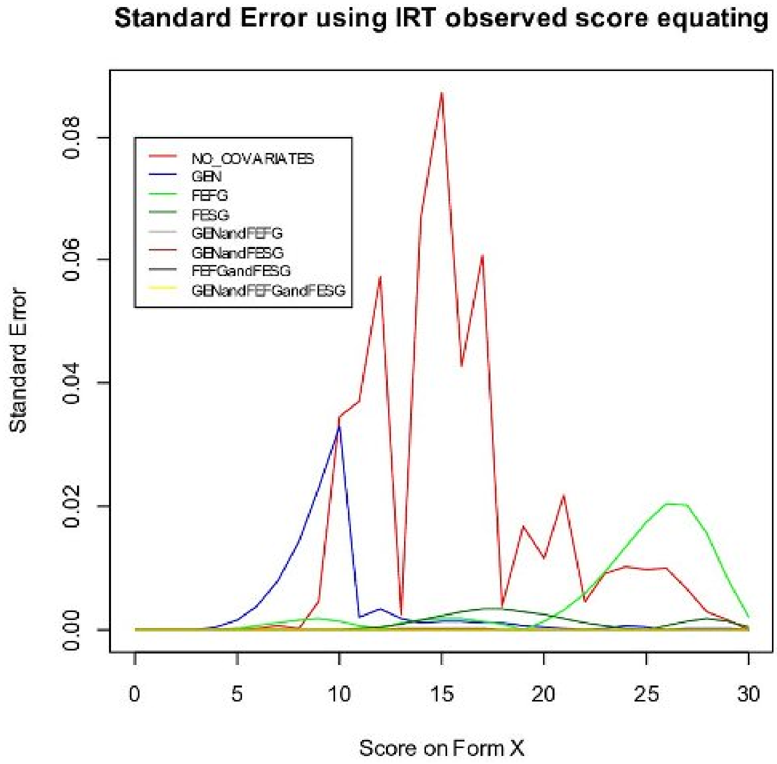


Figure A.23: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 600$ and DIF is low.

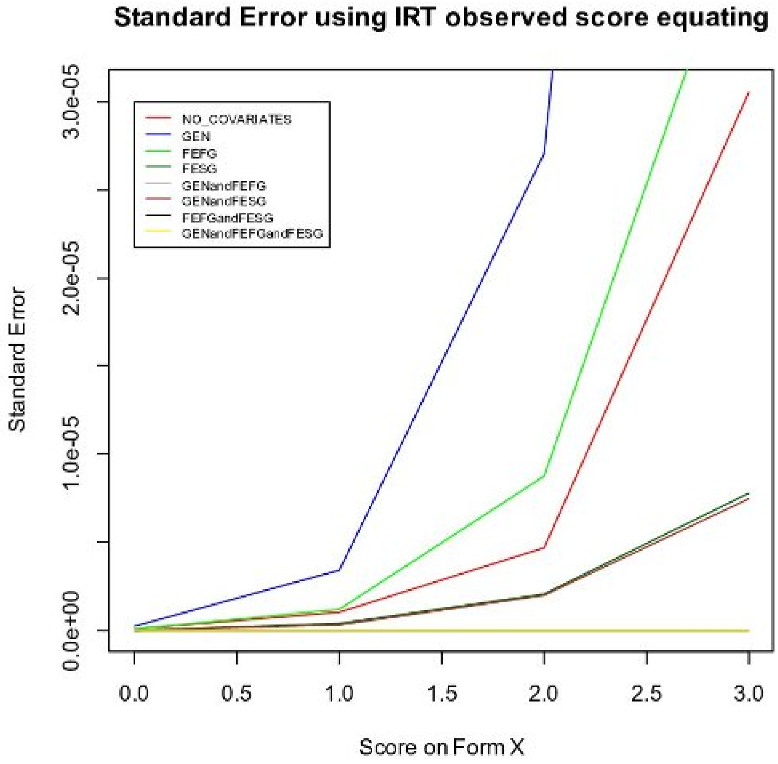


Figure A.24: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 600$ and DIF is low.

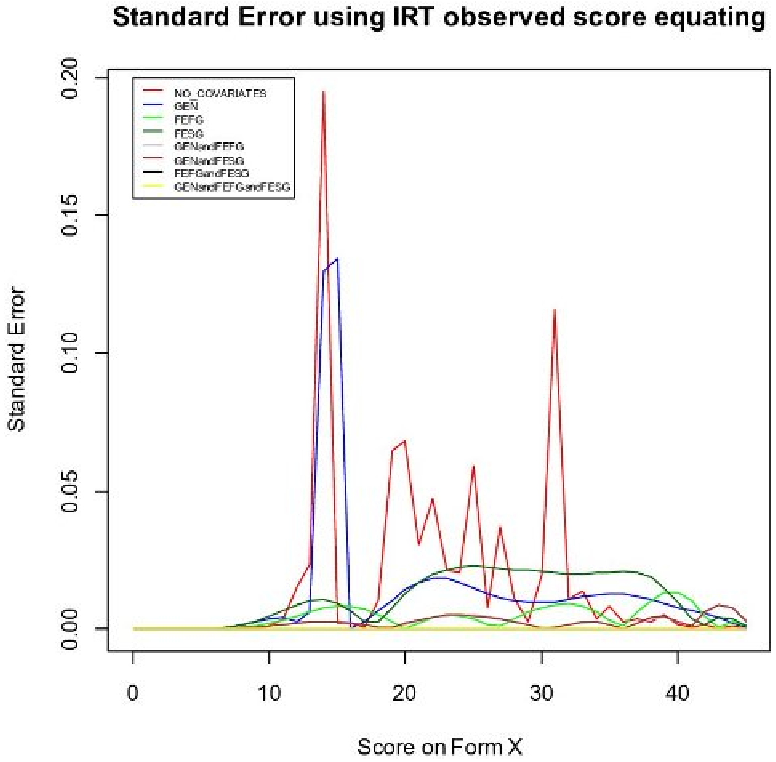


Figure A.25: SSE using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is low.

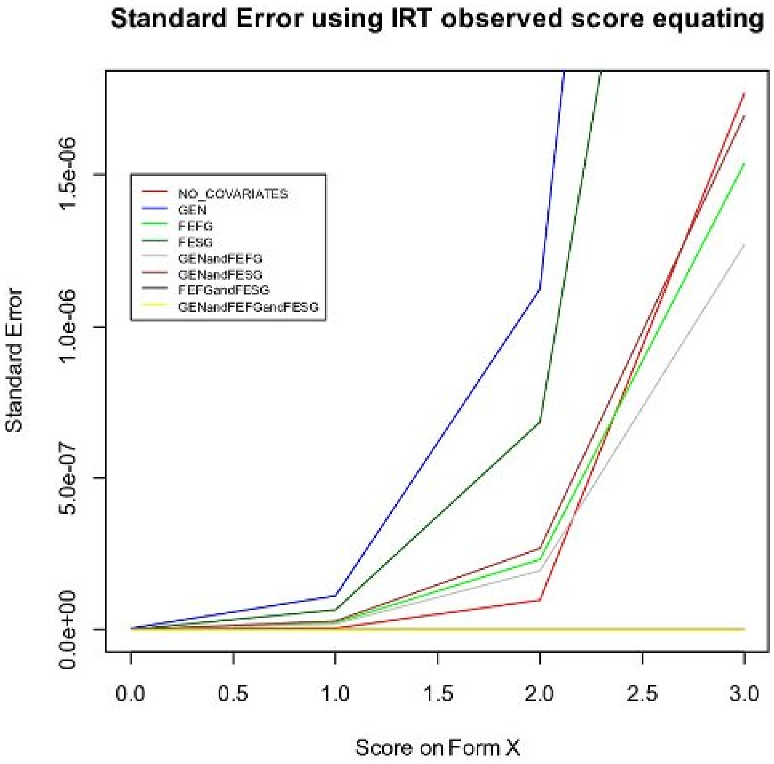


Figure A.26: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 45$, $N1 = N2 = 2000$ and DIF is low.

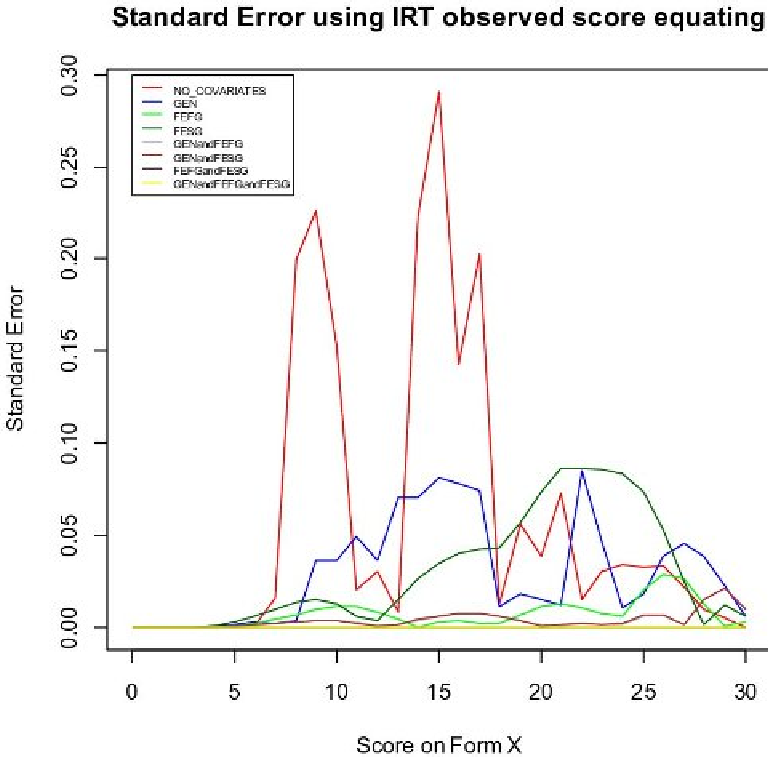


Figure A.27: SSE using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 2000$ and DIF is low.

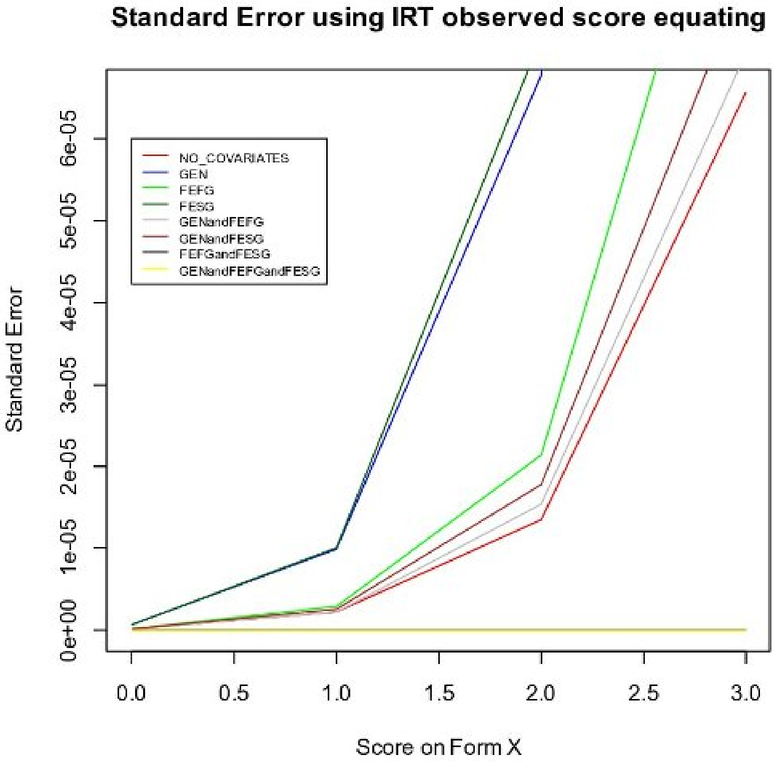


Figure A.28: SSE at the lower end of the scale using IRT observed-score equating when $n1 = n2 = 30$, $N1 = N2 = 2000$ and DIF is low.