HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Educational Sciences
Educational Measurement and Evaluation Program

COMPARISON OF EQUATING METHODS FOR MULTIDIMENSIONAL TESTS WHICH CONTAIN ITEMS WITH DIFFERENTIAL ITEM FUNCTIONING

Seçil UĞURLU

Ph.D. Dissertation

Ankara, 2020

With leadership, research, innovation, high quality education and change,

To the leading edge... Toward being the best...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Educational Sciences
Educational Measurement and Evaluation Program

COMPARISON OF EQUATING METHODS FOR MULTIDIMENSIONAL TESTS WHICH CONTAIN ITEMS WITH DIFFERENTIAL ITEM FUNCTIONING

DEĞİŞEN MADDE FONKSİYONU GÖSTEREN MADDELER İÇEREN ÇOK BOYUTLU TESTLERİN EŞİTLENMESİNDE KULLANILAN YÖNTEMLERİN KARŞILAŞTIRILMASI

Seçil UĞURLU

Ph.D. Dissertation

Ankara, 2020

Abstract

Differential item functioning (DIF) and test equating are two important subjects for fairness in reported test scores. However, they have been treated separately in the psychometric literature. Hence, studies investigating the relationship between DIF and equating are quite rare. Moreover, there is no study addressing this relationship in multidimensional perspective, yet. The purpose of this study is to reveal the relationship between DIF and equating in multidimensional perspective, and contribute to the literature. In order to reveal this relationship accurately and clearly, population invariance of equating was investigated. The data used in the study were generated in accordance with the simple structure multidimensional item response theory (SS-MIRT) model. Four different equating methods were used in the study: simple structure-multidimensional item response theory observed score, unidimensional item response theory observed score, unidimensional item response theory true score and equipercentile equating. The performances of the equating methods were compared according to their population invariance under various simulation conditions (differential form DIF, correlation between dimensions, group mean ability differences between two forms). According to the results, the method that most accurately reflects the relationship between DIF and equating is the multidimensional equating method for the 0.5 correlation condition. On the other hand, under 0.8 and 0.95 correlation conditions, all methods give similar results except the results of equipercentile equating method at scores with low frequencies. Also, group mean ability difference between two forms has no effect on the population invariance of the methods.

Keywords: differential item functioning, DIF, test equating, multidimensional item response theory, multidimensional test equating, population invariance, equating invariance.

Öz

Değişen madde fonksiyonu ve test eşitleme çalışmaları test puanlarının adil değerlendirilmesine hizmet etmektedirler. Ancak bu iki kavram psikometrik literatürde çoğunlukla ayrı ayrı ele alınmıştır. Literatürde değişen madde fonksiyonu ve eşitleme arasındaki ilişkiyi araştıran çalışma sayısı oldukça azdır. Dahası bunu çok boyutlu perspektifte ele alan çalışmaya henüz rastlanmamıştır. Literatürdeki bu eksikliği gidermeye yönelik olarak bu araştırmanın amacı değişen madde fonksiyonu ve eşitleme arasındaki ilişkiyi çok boyutlu perspektifte ortaya koymaktır. Bu ilişkiyi doğru ve açık bir şekilde ortaya koyabilmek için eşitlemenin popülasyon değişmezliği özelliği araştırılmıştır. Çalışmada kullanılan veriler basit yapılı-çok boyutlu madde tepki kuramına uygun olacak şekilde üretilmiştir. Çalışmada dört farklı eşitleme yöntemi kullanılmıştır: basit yapılı-çok boyutlu madde tepki kuramı gözlenen puan, tek boyutlu madde tepki kuramı gözlenen puan, tek boyutlu madde tepki kuramı gerçek puan ve eşit yüzdelikli eşitleme. Çalışmada kullanılan eşitleme yöntemlerinin performansları çeşitli simülasyon koşulları altında (formlar arası farklılaşan değişen madde fonksiyonu, boyutlar arası korelasyon, formlar arası grup yetenek ortalamaları farkı) popülasyon değişmezlik sonuçlarına göre karşılaştırılmıştır. Araştırmanın sonuçlarına göre değişen madde fonksiyonu ve eşitleme arasındaki ilişkiyi en doğru şekilde yansıtan yöntem 0.5 korelasyon koşulu için çok boyutlu eşitleme yöntemidir. 0.8 ve 0.95 korelasyon koşullarında ise, eşit yüzdelikli eşitleme yönteminin düşük frekanslı puanlarda verdiği sonuçlar dışında tüm yöntemler benzer sonuçlar vermektedir. Formlar arasındaki grup yetenek ortalaması farkının ise yöntemlerin popülasyon değişmezlik sonuçları üzerinde bir etkisi bulunmamaktadır.

Anahtar sözcükler: değişen madde fonksiyonu, DMF, test eşitleme, çok boyutlu madde tepki kuramı, çok boyutlu test eşitleme, popülasyon değişmezliği, eşitleme değişmezliği.

To my little daughter Doğa, and my dear husband Bilal

Acknowledgements

I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. Burcu Atar for her encouragements, guidance, advices, criticism and insight throughout the research.

I would like to extend my thanks to Prof. Dr. Selahattin Gelbal for his constant help, advice, and encouragement during my entire Ph.D. education.

I sincerely thank my other committee members, Prof. Dr. Nuri Doğan, Prof. Dr. Hakan Yavuz Atar, and Prof. Dr. Cem Oktay Güzeller for their comments to improve my dissertation.

I would also like to express my fullest appreciation to Prof. Dr. Won-Chan Lee for his suggestions, comments, and invaluable helps.

I would also like to express my sincere appreciation to my colleague and friend Dr. Emny Sousa-Bernini for reading my dissertation.

I want to express my thanks to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for their financial support during my dissertation research.

I would like to express my very special thanks to my family and my husband's family for their patience, understanding, and morale support.

My deepest feelings of gratitude go to my lovely daughter Doğa. Having her near me gave me strength and hope. My last but not least gratitude goes to my husband Bilal UĞURLU for his endless love and patience. He has always been there for me whenever I needed him. Without his patience, love and encouragement, I would have never been this determined.

Table of Contents

List of Figures

Symbols and Abbreviations

DIF: Differential item functioning

DTF: Differential test functioning

IRT: Item response theory

UIRT: Unidimensional item response theory

CINEG: Common-item nonequivalent groups

CTT: Classical test theory

MIRT: Multidimensional item response theory

EQ: Equipercentile

RMSD: Root mean square difference

REMSD: Root expected mean square difference

RESD: Root expected squared difference

RSD: Root squared difference

RG: Random groups

SG: Single group

PL: Parameter logistic

3-PL: Three-parameter logistic

ICC: Item characteristic curve

TCC: Test characteristic curve

ICS: Item characteristic surface

TCS: Test characteristic surface

SS-MIRT: Simple structure-multidimensional item response theory

UO: Unidimensional item response theory observed score

UT: Unidimensional item response theory true score

SMO: Simple structure-multidimensional item response theory observed score

SMT: Simple structure-multidimensional item response theory true score

MC: Multiple choice

FR: Free response

CE: Chained equipercentile

BF-MIRT: Bi-factor multidimensional item response theory

TRT: Testlet response theory

FMIRT: Full multidimensional item response theory

SL: Stocking-Lord

ES: Effect size

ETS: Educational Testing Service

R: Reference

F: Focal

BN: Bivariate normal

TRM-MIRT: Testlet response model- multidimensional item response theory

DTM: Difference that matters

Chapter 1
Introduction

Test scores are commonly used to provide information when making important decisions. Some of these decisions focus at the individual level, some focus at the institutional level, and some others focus at the public policy level. Making decisions in these contexts require tests to be applied on multiple occasions (i. e. test dates). If same test items are administered on each occasion, examinees may inform each other about the test items, or examinees who tested twice may remember the items. In both cases, security problems occur. To handle these problems, a different set of test items can be administered to the examinees who are tested on different test dates. This different set of test items is referred as a test form. There may be differences in difficulties of the test forms even if they are built to be similar in their content and difficulty.

Equating is a statistical process to adjust test scores on different test forms, and gives us the opportunity of using test scores interchangeably by adjusting scores for difficulty differences. It should be emphasized that scores are adjusted for form difficulty differences not for form content differences in an equating process. After a successful equating process, it can be said that the equated test scores have the same meaning. That is, for equated scores, it does not matter to whom and when the test was administered. To be able to conduct equating, test forms should be as similar as possible in content and statistical characteristics, otherwise this statistical process cannot go beyond linking. Although, linking and equating use similar statistical procedures, their purposes are different. Specifically, linking relate tests that are built to be different, while equating adjusts test scores on different forms that are built to be similar in content and statistical characteristics (Kolen & Brennan, 2014).

In order to conduct equating properly, five basic requirements should be met in addition to the techniques used in test linking: (1) same construct requirement, (2) equal reliability requirement, (3) symmetry requirement, (4) equity requirement, and (5) population invariance requirement. According to Dorans and Holland (2000) *"population invariance"* is one of the most important requirements of equating. This requirement means that equating function used to link scores obtained from test forms should be population invariant. That is, it does not matter which subpopulation

is used when equating test forms by using the equating function (Dorans & Holland, 2000). In other words, the function used to equate test scores should be the same for various populations or subpopulations. Otherwise, when equating functions of populations or subpopulations differing systematically, interchangeability of test scores is questionable (Kolen, 2004).

Dorans and Holland (2000) state that equating invariance is never exact because there are always some subpopulation effects on equating function. However, just for the test forms which are reliable and parallel in form and content, it may hold approximately. Flanagan (1951) also states that differences in content and reliability across forms can cause population dependence of equating. Green's (2003) assertion also confirms Flanagan's belief about the relationship between content, reliability and population dependence of equating. In a following study, Kolen (2004) indicates the same by referring equating theory and empirical research.

Lack of equating invariance means that difference between the difficulties of two forms is not consistent across subpopulations. In other words, there is an interaction among the relative difficulty of two forms and subpopulation membership. In this case, scores obtained from two forms are not interchangeable and hence treating subpopulations the same by using same conversions is a concern for fairness and score equity (Dorans, 2004, 2008).

Differential item functioning (DIF) is another facet of fairness. DIF occurs when probability of correct response to the same item for examinees at the same ability level differs according to group membership (Camilli & Shepard, 1994). Lack of DIF, or presence of test fairness in DIF perspective, requires the regression of item score onto the total score to be invariant across subpopulations. Hence, subpopulation invariance is also a key concept in a DIF study (Dorans, 2004). Although equating and DIF are both interested in subpopulation invariance, they have been treated separately in psychometric perspective (Huggins, 2012, 2014).

Item bias may occur when test scores are obtained as if they are measuring only a single ability but in fact are capable of measuring multiple abilities. Let us assume that there are two subpopulations having different underlying ability distributions and test items that are able to measure these multiple abilities. In such

a case, using unidimensional item response theory (UIRT) and getting test scores as if they reflect only one ability may cause item bias (Ackerman, 1992). In the case of multidimensional DIF, there are multiple traits relevant to the test's purpose and some test items are sensitive to irrelevant or nuisance traits (Mazor, Hambleton & Clauser, 1998). In fair test use, there are not any strict rules aiming to measure only one dimension; on the contrary, multidimensionality is very useful for fair test use but is confusing for DIF (Dorans, 2004). Because, as mentioned above, the multidimensional structure of the item increases the potential of existence of DIF. In other words, we need multidimensionality for fairness in assessment. However, it is very difficult to mention the lack of DIF in this case.

For common-item nonequivalent groups (CINEG) design, that is frequently used in actual applications and also used in this research, DIF in common items has the potential to result in systematic equating error and cause the lack of equating invariance (Kolen & Brennan, 2014). Hence, to represent the relationship between DIF and equating clearly and accurately, equating invariance should be investigated in addition to DIF analysis. DIF focuses on items whereas equating focuses on reported scores. Therefore, the assessment of DIF is not sufficient alone to represent the effect of DIF on reported scores and there is a need for assessment of population invariance of equating.

In most cases, there is a need for an examinee to have more than one ability to answer an item correctly. The necessity of having more than one ability is called multidimensionality. We need multidimensionality for fair assessment but the presence of multidimensionality has the potential to cause DIF and equating dependence. In most cases, practitioners use unidimensional methods to score and equate tests that are multidimensional. Currently, multidimensional methods are popular and getting more attention in scoring and equating (Lee & Brossman, 2012; Lee & Lee, 2016; Lee, Lee & Brennan, 2014; Kim, Lee & Kolen, 2020; Peterson & Lee, 2014).

Aim and Significance of the Study

Standard tests used to measure student achievement mostly require multiple abilities. In this respect, data sets are mostly multidimensional. However, DIF occurs when items are related to an undesired dimension other than the desired ones.

Therefore, it is unlikely to talk about lack of DIF in the presence of multidimensional data. Hence, examining the results of equating conducted with multidimensional data under the presence of DIF is an important research topic. In this respect, the results of multidimensional equating conducted in the presence of DIF are aimed to be discussed in this study. In other words, the goal of this study is to reveal the effects of DIF on equating in a multidimensional perspective. On the other hand, in the literature, generally unidimensional equating methods have been used to equate multidimensional data sets since the multidimensional equating methods are quite new and complex to use. However, it is recommended (Lee & Brossman, 2012; Lee & Lee, 2016; Lee, Lee & Brennan, 2014; Kim, Lee & Kolen, 2020; Peterson & Lee, 2014) that equating studies conducted with multidimensional data sets should be based on multidimensional equating methods to reduce the equating error. In the current study, it is also aimed to compare the performance of multidimensional and unidimensional equating methods in the presence of multidimensional data structure based on the population invariance of the methods. Thus, the performances of multidimensional equating methods are discussed in terms of population invariance, and this will make an important contribution to the psychometric literature.

According to the literature review, researches that reveal the relationship between DIF and equating are quite rare. One of the most basic studies was conducted by Dorans (2004) in Classical Test Theory (CTT) perspective. However, this research includes theoretical information only. Hence, in this study, it is aimed to prove this theoretical knowledge on different applications. Another basic study aimed at dealing with the relationship between DIF and equating invariance was conducted in Item Response Theory (IRT) perspective and belongs to Huggins (2014). Unlike this study, the relationship between DIF and equating is aimed to be examined in a multidimensional perspective in the current study. There are also some other studies in which changes were observed in equating errors for conditions where DIF exists (Atalay Kabasakal & Kelecioglu, 2015; Demirus & Gelbal, 2016; Yurtcu & Guzeller, 2018). However, to reveal the relationship between DIF and equating accurately and clearly, population invariance should be examined in detail as suggested by Dorans (2004), and Huggins (2014). From this point of view, in the current study population invariance is emphasized. On the other hand,

multidimensional item response theory (MIRT) equating is an essential study subject. However, there is a limited number of studies conducted on this subject yet (Lee & Brossman, 2012; Lee & Lee, 2016; Lee, Lee & Brennan, 2014; Kim, Lee & Kolen, 2020; Kumlu, 2019; Peterson & Lee, 2014). Moreover, no research has appeared so far regarding the relationship between DIF and multidimensional equating. Therefore, it is quite important to discuss this relationship in MIRT perspective. Also, with the current study, it is aimed to reveal which equating methods serve to fair assessment when the data is multidimensional and there are some DIF items in the common item set. And, it is also aimed to show what kind of drawbacks may arise when an equating method that is inappropriate for a specific condition is used.

Research Questions

Under the light of the information given above, the goal of this study can be clearly stated as to investigate the effect of differential item functioning on population invariance of multidimensional IRT (MIRT), unidimensional IRT (UIRT), and Equipercentile (EQ) equating methods, when equating multidimensional test forms that contain DIF items in common item set.

Below are the specific research questions:

1. What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of differential form DIF on population invariance?
2. What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of correlation between dimensions on population invariance?
3. What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of group mean ability differences between two forms on the relationship between DIF and population invariance?

Definitions

Another point to be mentioned is the definitions of "no-DIF, DIF in both forms, and DIF in new form only", which are frequently used in the research. Accordingly, “no-DIF” refers to the absence of DIF in both forms. “DIF in both forms” states that the same amount of DIF is added in the same direction to the same common items of two forms. And, “DIF in new form only” states that while there is no DIF in the common items in the old form, DIF is added to the common items in the new form. Thus, the difficulties of the common items in the two forms are differentiated from each other. And accordingly, these three conditions are collectively defined as "differential form DIF".

Limitations of the Study

Before proceeding with the literature review, it is desired to mention about the time limitation, which was an important limitation of this research. Parameter estimation of multidimensional data with concurrent calibration in this process required a very long time. Therefore, the research had to be limited to 100 iterations.

On the other hand, existence of DIF in common items is more problematic compared to existence in non-common items in terms of equating invariance results. Therefore, in this study, DIF was added only to common items to examine the effect of DIF on equating invariance. It can be said that the presence of DIF in common items is also a limitation for this research, as the equating is conducted by using the common items in CINEG design.

Chapter 2
Literature Review

This chapter provides a general overview of linking, equating, equating methods, equating designs, population invariance of equating, differential item functioning (DIF), relationship between equating invariance and DIF, and some other important concepts. At the end of this chapter, a detailed literature review is also provided.

Linking and Equating

Tests are generally used to give information in making important decisions. These decisions may focus at the individual level, institutional level, or public policy level. Making decisions in these contexts requires tests to be applied on multiple occasions. If the same test questions are administered on each occasion (test date), some security problems may occur: examinees may inform each other about test questions, or examinees who tested twice may remember the questions. To prevent these problems, a different collection of test questions can be administered to the examinees, who are tested on different test dates. This different collection of test questions is referred as a test form. Although test forms are built to be similar in their content and difficulty, there may be some difficulty differences between them. Equating is a statistical process to adjust scores for these difficulty differences on different test forms, and hence gives us the opportunity of using test scores interchangeably. It should be noted that equating adjusts scores for form difficulty differences not for form content differences. After a successful equating, equated test scores have the same meaning. That is, for equated scores it does not matter to whom and when the test was administered. To be able to conduct equating, test forms should be as similar as possible in content and statistical characteristics otherwise this statistical process cannot go beyond linking. Although, linking and equating use similar statistical procedures, their purposes are different. Specifically, linking relate tests that are built to be different, while equating adjusts test scores on different forms that are built to be similar in content and statistical characteristics (Kolen & Brennan, 2014).

In order to conduct equating appropriately, five basic requirements should be met in addition to the techniques used in test linking: (a) The Same Construct

Requirement, (b) The Equal Reliability Requirement, (c) The Symmetry Requirement, (d) The Equity Requirement, (e) Population Invariance Requirement. According to population invariance requirement, the choice of (sub) population used to compute the equating function between the scores of two forms should not matter. In other words, the equating function used to link the scores of old and new forms should be population invariant (Dorans & Holland, 2000). This study focuses on the population invariance requirement of equating. Because, population invariance results are needed to reveal the relationship between DIF and equating, accurately. This issue is discussed in detail in the following section.

Equating Invariance

The primary purpose of equating is to use test scores, which are obtained from different forms interchangeably (Kolen, 2004). To be able to use test scores from alternate forms interchangeably, the equating function used to link their scores should be invariant across subpopulations (Dorans, 2004; Kolen, 2004). On the contrary, when the equating functions of populations or subpopulations differ systematically, it is questionable to use test forms interchangeably because it might result in a disadvantage for some subpopulations (Kolen, 2004; Powers & Kolen, 2014).

According to a large number of researches, it is not possible for an equating function to be completely population invariant. This is due to the constant existence of an effect on equating functions caused by important subpopulations. Instead, population invariance may hold approximately for tests that are parallel in form and content and are very reliable. Since, for test scores which suit these conditions the subpopulation does not have a very important effect on equating. In conclusion, subpopulation invariance never really holds exactly, but under proper conditions it may hold to some degree which is accepted to be sufficient (Dorans, 2004; Dorans & Holland, 2000; Petersen, 2008).

Besides, population invariance other important criteria must also be achieved to guarantee that a linking is an equating. According to Dorans and Holland (2000), there are five basic requirements of equating: (a) tests should measure the same constructs, (b) tests should have the equal reliability, (c) the equating function for equating the scores of Form 1 to the scores of Form 0 should be the inverse of the

equating function for equating the scores of Form 0 to the scores of Form 1, (d) it should not make a difference for an examinee to be tested by any of two tests that have been equated, (e) equating function used to link the scores of two tests should be invariant to subpopulations of examinees. The last requirement is population invariance which is the most important one for score equity. In assessing fairness, we need to assess subpopulation invariance. If subpopulation invariance does not hold to a sufficient degree, linking functions can still be computed but we cannot call this linkage as *equating* although the same statistical computations are carried out for both (Dorans, 2004; Dorans & Holland, 2000). It happens because, for saying that a linking is as an equating, test scores obtained from alternate forms should be interchangeable. In such a situation, the use of equated scores as if they were population invariant results in a big mistake in fair assessment. It would be a better idea not to conduct equating instead of causing an unfair assessment. To sum up, subpopulation invariance is very useful to assess fairness but of course it is not sufficient alone for equating. While lack of invariance indicates that a linking is not an equating, existence of invariance does not indicate that score interchangeability is achieved by equating (Dorans, Liu & Hammond, 2008).

Examining population invariance of equating is found useful by many researchers for all testing programs, which are interested in high-stakes outcomes (Brennan, 2008; Dorans, 2004; Kolen, 2004; Petersen, 2008). Assessing equating invariance is termed as score equity assessment by Dorans (2004, 2008) and in addition to DIF analysis, score equity assessment is expressed as another facet of fairness. Both of them use population invariance to assess fairness, however DIF is interested in items while equating is based on test scores. Therefore, using only DIF analyses is not sufficient to assess equated scores. As pointed out by Dorans (2004), some testing programs only address DIF, not score equity assessment. With this study, DIF and equating invariance aspects that are treated separately in psychometric perspective are discussed together. Besides, these aspects are tried to be expressed in multidimensionality perspective, which is one of the most interesting research topics recently. However, before talking about IRT methods, the methods used to measure equating invariance will be mentioned. These methods and their classification are as follows.

Methods for evaluating equating invariance. There are multiple methods for evaluating equating invariance in the psychometric literature. A taxonomy categorizing available methods were presented by Huggins and Penfield (2012). Based on this, methods were categorized into two sections. In the first section, the focus is how between-group differences are aggregated or separated across specific subpopulations. This section covers three different approaches: omnibus, group-to-overall, and pairwise. In the next section of current dissertation, omnibus and group-to-overall methods were explained.

Omnibus methods. In this approach, there is a single index that computes the distance between each subpopulation's linking function and the overall (all groups) linking function. For example, to compute the degree of equating invariance with respect to gender, three equating functions are estimated (i.e., one for males, one for females, and one for overall), and then a chosen omnibus method compares both male and female functions simultaneously to the overall equating function. Here, the computed index represents the equating invariance or lack of equating invariance (i.e., equating dependence) (Huggins, 2012; Huggins & Penfield, 2012).

Equating invariance methods can be conditional or unconditional on score level. When an omnibus method measures the invariance of equating relationships between subgroups and the overall group at each score level, this method belongs to the conditional omnibus method section in the taxonomy. On the other hand, when an omnibus method measures the same thing but this time across all score levels, this method belongs to the unconditional method section in the taxonomy (Huggins, 2012; Huggins & Penfield, 2012).

One of the most used omnibus methods in the psychometric literature is the root mean square difference (RMSD) (Dorans & Holland, 2000) which is conditional on score levels. In current dissertation the RMSD index was used to estimate the equating invariance of various equating methods. This index, which is appropriate to be used for single or equivalent group design and linear linking function, can be explained in detail as below:

$$RMSD(x) = \frac{\sqrt{\sum_j w_j \left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{YP}}$$

The above equation is computed at each x value. Where, j represents different subgroups, P represents the overall group, $e_{P_j}(x)$ represents separate linking functions for specified subgroups, $e_P(x)$ represents the overall linking function, w_j represents the weighting of each subgroup (a proportional representation of each subgroup in the overall group), σ_{YP} represents the standard deviation of Y scores in P (Dorans & Holland, 2000).

According to some studies (von Davier, Holland & Thayer, 2004; von Davier & Wilson, 2008), RMSD index can be adopted to other equating designs and methods. In current study, unstandardized version of this index (obtained by removing the dominator component) was used to equate test forms with CINEG design.

Root expected mean square difference (REMSD) (Dorans & Holland, 2000), which is another omnibus method, is an unconditional version of RMSD index. Dorans and Holland (2000) created this index to summarize the values of RMSD (x) into a single number. This index is an average over the distribution of X in P. The REMSD (x) is as below:

$$\text{REMSD} = \frac{\sqrt{E_P\left\{\sum_j w_j\left[e_{P_j}(X) - e_P(X)\right]^2\right\}}}{\sigma_{YP}} = \frac{\sqrt{\sum_j w_j E_P\left\{\left[e_{P_j}(X) - e_P(X)\right]^2\right\}}}{\sigma_{YP}}$$

where E_P represents averaging over the distribution of X in P.

Group to overall methods. Group to overall invariance methods compares one subgroup's linking function to the overall linking function ignoring all other subgroups' linking functions. These methods produce a separate index of equating invariance for each subgroup (Huggins & Penfield, 2012). One of the group-to-overall invariance indices is the root expected squared difference (RESD_j) (Yang, 2004). RESD_j which is an unconditional index is explained below (von Davier & Liu, 2006; Yang, 2004):

$$RESD_j = \frac{\sqrt{E_P\left\{\left[e_{P_j}(x) - e_P(x)\right]^2\right\}}}{\sigma_{Y_P}} = \frac{\sqrt{\sum_{x=0}^{Z} w_{x_P}\left\{\left[e_{P_j}(x) - e_P(x)\right]^2\right\}}}{\sigma_{Y_P}}$$

where j represents a subgroup, w_{x_P} represents the weighting of the relative number of candidates in the total population, $\sum_{x=0}^{Z}$ represents the averaging the weighted differences at each score level, σ_{Y_P} represents the standard deviation of the composite score in the total population, and the others are the same with the above equations.

Root squared difference ($\mathrm{RSD}_j(x)$), which is a conditional version of $RESD_j$ invariance index, is created by Huggins and Penfield (2012). This index is presented below:

$$RSD_j(x) = \frac{|d_j(x)|}{\sigma_Q}$$

where σ_Q represents the standard deviation of scores in population Q, and $d_j(x)$ represents the difference between a linked score y based on subgroup j's linking function and a linked score y based on the overall linking function at score level x (Huggins & Penfield, 2012). This index gives the standardized distance between the one subgroup's equating function and the overall equating function at one score level (Huggins, 2012).

Above, the methods for evaluating equating invariance are discussed in detail. Another important topic to be addressed in this study is the equating designs. Hence, detailed information about the equating designs frequently used in the literature is given below.

Equating Designs

In psychometric literature, there are various equating designs to conduct test equating. In the literature, three most commonly used equating designs are: random groups (RG), single group (SG), and common-item nonequivalent groups (CINEG) designs.

In RG design, examinees are randomly assigned to different test forms. Difference between group level performances on two forms is attributed to the difficulty difference between two forms, because examinee groups who take different forms are considered to be equivalent in their ability level. One advantage of this design is that each examinee takes only one form and this feature minimize testing time compared to designs that request examinees to take more than one

form. On the other hand, this design requires all test forms be administered at the same time. And also, large sample sizes are needed because different examinees take different forms. These are the limitations of this design.

In SG design, the same examinees take both forms, Form X and Form Y. Because of the possible effects of fatigue and familiarity with the test on the performance, this design has some strong limitations. Specifically, for examinees, taking two forms consecutively may cause fatigue. In such a situation, Form Y (second form) could appear relatively more difficult than Form X. On the other hand, there may be familiarity with the forms. In this case, Form Y could appear to be easier than Form X. Counterbalancing the order of administration of the forms is a common method to deal with these undesirable effects. However, even so this design is rarely used in practice.

The last design is called common-item nonequivalent groups (CINEG) design in this dissertation. This design has been referred to as common-item nonequivalent groups design by some researchers, while it has been referred to as non-equivalent anchor test design by some others. CINEG is preferred in this study. According to CINEG design, there are two different examinee groups and they take two forms, which have a set of common items. Two forms are linked by using these common items. It should be noted that common items should be the representative of total test forms in their content and statistical characteristics, like a mini version of total tests. Another important point is that, the examinee groups are nonequivalent. Hence, group differences should be separated from form differences. Finally, it is a very popular design because of the similarity with the actual testing situations (Kolen & Brennan, 2014).

It can be said that CINEG design is a complex design to execute well because of the differences in ability between the old and new groups. Complexity increases as the differences increase. Also, the type of common item set, external or internal, and type of score linking also have an effect on the complexity of this design (Dorans, Moses & Eignor, 2010).

Another important topic of this study is the equating methods. In this study MIRT, UIRT and EQ equating methods are used. Before mentioning these methods, MIRT and UIRT models will be discussed in detail. These models are used in many

testing situations. There are many IRT applications in the literature, such as test development, item banking, differential item functioning, adaptive testing, test equating, and test scaling. IRT models examine responses at the item level, whereas, classical test models examine responses at the level of test scores. This makes the IRT model very powerful. In this study IRT models are used to generate data, and then IRT and the traditional EQ equating method results are compared.

Item Response Theory

IRT can be classified in two sections according to the number of underlying latent traits in the model: UIRT and MIRT. In UIRT there is only one underlying latent ability that is specified in the model whereas in MIRT this is more than one.

Unidimensional item response theory. In unidimensional item response theory, item response function represents a mathematical statement as to how response depends on level of ability or skill. For a dichotomous item, the item response function is simply the probability (P or P (θ)) of a correct response to the item. Based on a common assumption this probability can be represented by the logistic function (Lord, 1980). Three parameter logistic function (Birnbaum, 1968), which is the most popular in logistic models, can be expressed as below:

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-1.7\alpha_j(\theta_i - b_j)}},$$

where

θ_i is the examinee i ability parameter,

α_j is the item j discrimination parameter,

b_j is the item j difficulty parameter,

c_j is the item j pseudo-guessing or lower asymptote parameter,

1.7 is the scaling factor *D*, which is used to make the logistic function as close as possible to the normal ogive function. When c_j equals to zero, three parameter logistic (PL) model turns into two PL model. And, when α_j equals to one, two PL model turns into one PL model (Hambleton, Swaminathan & Rogers, 1991; Lord, 1980). The number of item parameters that is used in the logistic model determines

the type of the model: one, two and three parameter logistic models. There is also a four-parameter logistic model in the literature.

Item response function can be represented graphically by the item characteristic curve (ICC). On the other hand, test characteristic curve (TCC) graphically represents the test characteristic function that gives the relationship of the number right score and ability level (Lord, 1980). In other words, TCC is used to represent the characteristics of a test and can be found by summing the item characteristic curves across all items in the test. That is, this curve is the regression of the sum of the item scores on θ. For the sake of clarity, visual examples of the ICC and TCC are shown in Figure 1 and Figure 2, respectively.

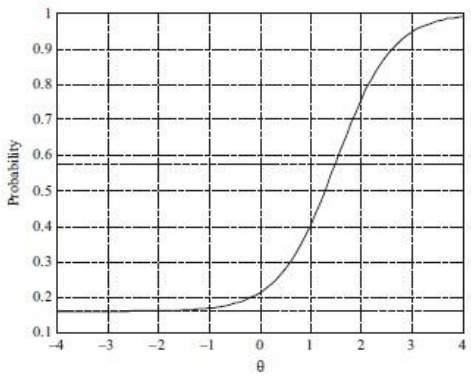

Figure 1. ICC for an item described by a three PL model (Reckase, 2009).

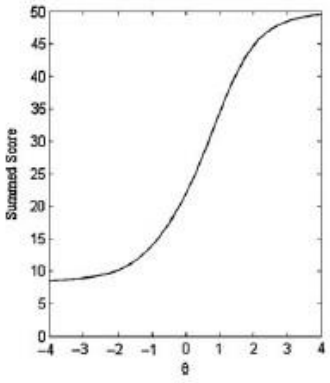

Figure 2. TCC for items described by a three PL model (Reckase, 2009).

This section provides some basic notions about UIRT. In order to use this model effectively, some assumptions should be met. Unidimensionality, monotonicity, and local independence are some of them. When unidimensionality

assumption holds, there are only one ability in the latent space. On the other hand, the monotonicity assumption means that the probability of correct response to the item increases, or at least does not decrease, as the ability level increase. Finally, according to the local independence assumption, the items should only be correlated through the latent trait that the test is measuring, should not be related to each other. However, things are more complicated than thought and it is hard to meet these assumptions. And hence, a more complicated model is needed to explain individuals' responses. This is where MIRT models emerge.

According to the literature, research subjects of UIRT have been generalized to MIRT models. Differential item functioning (Ackerman, 1992; Li & Stout, 1996; Shealy & Stout, 1993), equating (Lee & Brossman, 2012; Lee & Lee, 2016; Kim, Lee & Kolen, 2020), and scale linking (Davey, Oshima & Lee, 1996; Li & Lissitz, 2000; Oshima, Davey & Lee, 2000; Yao, 2011; Yao & Boughton, 2009) are the examples of these research subjects that are being studied in multidimensional perspective. MIRT models are discussed in more detail below.

Multidimensional item response theory. Unidimensionality, which requires a single ability parameter, is a strong assumption of IRT. In actual testing situations, it is difficult to meet this assumption. It occurs because examinees are likely to have more than one ability to answer an item correctly, especially for natural sciences. Hence, although UIRT is very useful under certain conditions, there is a need for MIRT models, which are more complex and capable of reflecting the complex relationship between examinees and test items more accurately (Kolen & Brennan, 2014; Reckase, 2009).

In MIRT models, there is a complex multidimensional space to describe individual differences in the target traits. A general representation of a MIRT model is represented as follows:

$$P_i(U = u|\theta) = f(\theta, \eta_i, u).$$

In this equation, θ, which contains *m* ability parameters ($\theta = (\theta_1, \theta_2, \ldots, \theta_m,)$) for the m dimensional space, is a vector of person parameters, η is a vector of item parameters, U is the score on the item, and u is a possible value for the score. It should be noted that in this mathematical representation the item score, u, exists only when the item has two score categories (Reckase, 2009).

In multidimensional item response theory, item response function can be represented graphically by the item characteristic surface (ICS). On the other hand, test characteristic surface (TCS), which can be found by summing the item characteristic surfaces across all the items in the test, is used to represent the characteristics of a test. That is, test characteristic surface is the regression of the sum of the item scores on θ vector (Reckase, 2009). Visual examples of the ICS and TCS are shown in Figure 3 (Reckase, 2009) and Figure 4, respectively (Reckase, 2007).

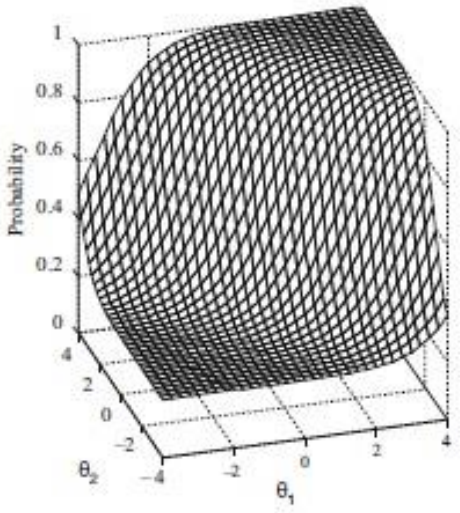

Figure 3. ICS for an item with $a_1 = 1.3, a_2 = 1.4, d = -1, c = 0.2$ (Reckase, 2009).

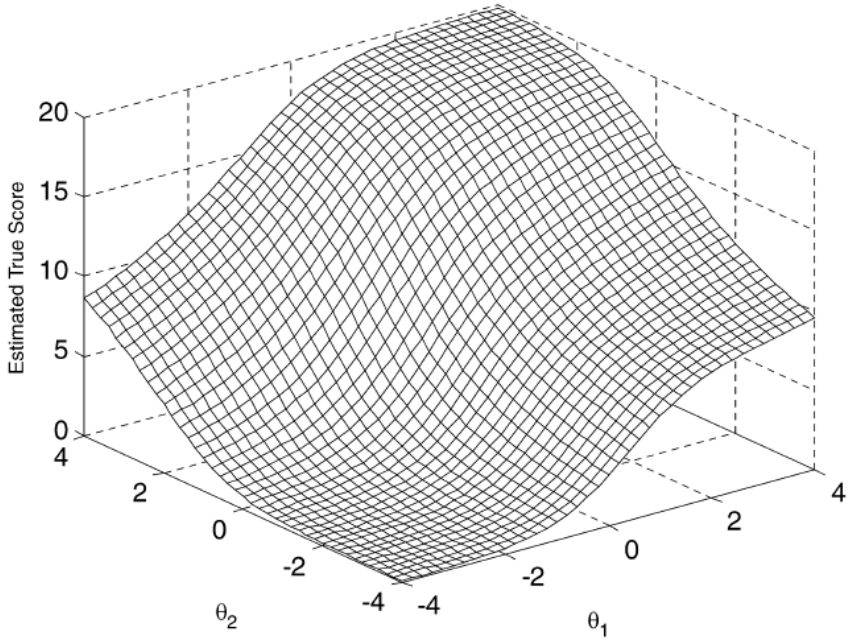

Figure 4. TCS for a 20-item test (Reckase, 2007).

Monotonicity and local independence are two basic assumptions of MIRT models. Monotonicity assumes that the probability of a correct response to an item increases with an increase of any element in the θ vector. Local independence assumes that all examinees respond all test items independently. That is, the response of any person to any item only depends on the vector of person parameters, θ, and the vector of item parameters, η (Reckase, 2009).

MIRT models can be classified in different ways. The two most basic of these are compensatory (Reckase, 1985) and non-compensatory (Sympson, 1978) models. Compensatory model is based on a linear combination of θ coordinates. Here in this type of MIRT model, one θ coordinate can compensate the other θ coordinate. So, the sum can be the same for various combinations of θ coordinates. So, according to this model an examinee with a low ability on one dimension will be able to answer the item correctly even if she/he has high abilities on other dimensions. The second type of model, which is called non-compensatory or partially compensatory model, uses a UIRT model for each dimension. Moreover, the probability of a correct response to the item is computed by the product of the probabilities for each part. This class of model results in a nonlinear feature because of the product of probabilities (Reckase, 2009). So, according to this model an examinee with a low ability on one dimension will not be able to answer the item correctly even if she/he has high abilities on other dimensions.

Another classification of MIRT models is based on the number of possible score points of the test items: MIRT models with dichotomously scored items (Bock & Aitken, 1981), and MIRT models with polytomously scored items (Adams, Wilson & Wang, 1997; Kelderman & Rijkes, 1994; Yao & Schwarz, 2006).

MIRT models can be classified also based on the dimensional structure of the items in the test. First type of model is complex structure MIRT models, which assume that items in the test measure more than one ability simultaneously. Multidimensional extension of the Rasch model (Adams, Wilson & Wang, 1997), the multidimensional two parameter logistic model (McKinley & Reckase, 1982), the multidimensional three parameter logistic model (Reckase, 1985), and the multidimensional two parameter partial credit model (Yao & Schwarz, 2006) are some of the examples of complex structure MIRT models. The second type of model is constrained MIRT models. According to this model, items in the test measure only

one or two abilities. Simple structure-MIRT (SS-MIRT) model (Segall, 1996) and bi-factor model (Gibbons et al., 2007; Gibbons & Hedeker, 1992) are the examples of constrained MIRT models. Comparing the complex structure and the constrained MIRT models, the constrained MIRT model has some advantageous over the complex structure MIRT model, because of its easier interpretability.

Simple structure MIRT models. According to Thurstone's (1947) principle of simple structure, for a given item the factor loadings should be relatively large, suggesting a clear relationship between the latent ability and item, or should be relatively small, suggesting no relationship between the latent ability and item (Finch, 2006; McLeod, Swygert & Thissen, 2001). In other words, latent traits have high loadings on some of the items when they have low loadings (close to 0) on the rest of the items (Finch, 2006).

A figure was added below to explain simple structure clearly. In Figure 5, there are ten items and two latent abilities (θ_1, θ_2) with correlation ρ. The first ability, θ_1, has high loadings on the first six items and the second ability, θ_2, has high loadings on the rest four items. Furthermore, there is a correlation between these two abilities. Without the correlation between abilities, this model reflects two separate UIRT models.

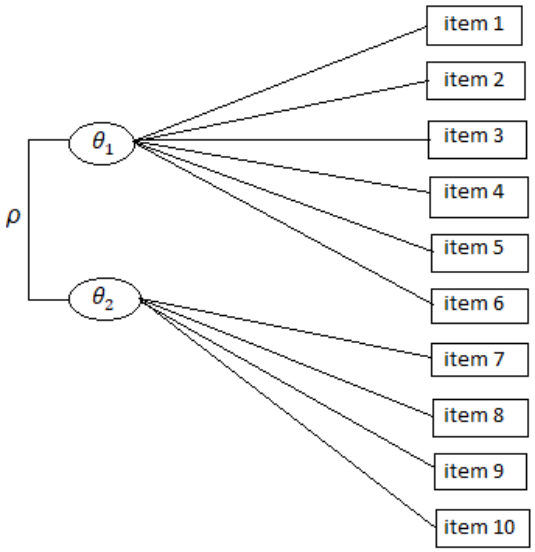

Figure 5. Simple structure-MIRT model.

There are some advantageous of SS-MIRT model compared to complex MIRT models. First, the number of item parameters does not change as the number of dimension increases. This feature gives the opportunity of reducing the estimation time of the analyses. Second, there is no need for an additional work to calibrate the item parameters, because of the multiple UIRT structure of this particular MIRT model. That is, estimation can be done by using UIRT item calibration for each separate data structure. Third, this model is very similar to UIRT in terms of general concepts and interpretation of parameters. Fourth, unlike other complex MIRT models, this model does not have a specific mathematical equation of item response function (IRF). Hence, combination of any kinds of UIRT models can be used to model the latent abilities (Kim, 2018). In conclusion, this model provides great flexibilities in many aspects and it has been used in many psychometric studies: calibration (Yao & Boughton, 2007; Zhang, 2004, 2012), equating (Kim, Lee & Kolen, 2020; Lee & Brossman, 2012), computerized adaptive testing (Li & Schafer, 2005; Luecht, 1996; Segall, 1996), and dimensionality (Li, Jiao, & Lissitz, 2012).

IRT Equating

Many testing programs use IRT models when creating their tests. Therefore, these programs generally use IRT equating methods when equating the tests. Also, IRT equating methods can be used for some specific situations in which traditional methods are not used (such as equating to an item pool). Thus, IRT methods are very important for equating methodology. However, IRT models have strong statistical assumptions. For this reason, for IRT applications it is crucial to investigate the violations of the assumptions, and IRT model fit (Kolen & Brennan, 2014). In this study IRT equating is examined under two headings: unidimensional and multidimensional IRT equating. First, UIRT equating is discussed as follows.

Unidimensional IRT equating. There are two IRT based equating methods in the literature: (1) IRT observed score, and (2) IRT true score equating methods (Kolen & Brennan, 2014).

Unidimensional IRT observed score equating. Unidimensional IRT observed score (UO) equating method obtains an estimated distribution of observed number correct scores for each form by using IRT models. Then, equating these distributions by using traditional EQ equating. First of all, it should be ensured that

item parameters are on the same metric. After that, for Form X conditional observed score distribution for examinees of a given ability is obtained by using the compound binominal distribution. A recursion formula (Lord & Wingersky, 1984) can be used to implement this process. This formula is given below:

$$\begin{aligned} f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1-p_{ir}), && x = 0 \\ &= f_{r-1}(x|\theta_i)(1-p_{ir}) + f_{r-1}(x-1|\theta_i)p_{ir}, && 0< x < r, \\ &= f_{r-1}(x-1|\theta_i)p_{ir}, && x = r \end{aligned}$$

where $f_r(x|\theta_i)$ is the probability of earning a score of x on the first rth item for examinee i with ability θ_i; and p_{ir} is the probability of a correct response on the rth item.

After the conditional distributions of observed scores for each ability level are obtained using the recursion formula above, the marginal observed score distribution is found by accumulating these observed score distributions using the equation below:

$$f(x) = \int_\theta f(x|\theta)\psi(\theta)d\theta,$$

where $\psi(\theta)$ is the distribution of θ. Because of the complexity of the integration, distribution of ability is approximated by a discrete distribution on a finite number of equally spaced points, which is referred as quadrature points. This approximation is as below:

$$f(x) = \sum_i f(x|\theta_i)\psi(\theta_i).$$

To conduct UO equating, the above steps should be done also for Form Y. As the final step, equating the two estimated marginal distributions is conducted by using traditional EQ equating methods (Kolen & Brennan, 2014).

Unidimensional IRT true score equating. According to unidimensional IRT true score (UT) equating, the true scores on Form X $T_X(\theta_i)$ and Form Y $T_Y(\theta_i)$ corresponding to a given θ_i are considered to be equivalent. This can be explained by the equation below,

$$irt_Y(T_X) = T_Y(T_X^{-1}), \qquad \sum_{j:X} c_j < T_X < K_X$$

where T_X^{-1} is the θ_i that corresponds to T_X true score. According to Kolen and Brennan (2014), UT equating consists of three steps:

1. Specify a true score T_X on Form X,
2. Find the ability level θ_i that corresponds to that true score (T_X^{-1}),
3. Find the true score T_Y that corresponds to the θ_i on Form Y.

Step 1 and 3 are straightforward but for step 2 an iterative procedure is needed. The Newton-Raphson method is used for finding the roots of nonlinear functions. In UT equating, to find the equivalent of T_X true score, first of all θ_i is going to be found by following the below equation:

$$func(\theta_i) = T_X - \sum_{j:X} P_{ij}\left(\theta_i; a_j, b_j, c_j\right)$$

and this function is going to be set to zero. The Newton-Raphson method uses the first derivative of $func(\theta_i)$ with respect to (θ_i). And hence, θ_i is tried to be find. The first derivative of $func(\theta_i)$ is as below:

$$func'(\theta_i) = -\sum_{j:X} P'_{ij}\left(\theta_i; a_j, b_j, c_j\right)$$

here $P'_{ij}(\theta_i; a_i, b_i, c_i)$ is defined as the first derivative of $P_{ij}\left(\theta_i; a_j, b_j, c_j\right)$ with respect to θ_i and is going to be calculated as below:

$$P'_{ij}(\theta_i; a_i, b_i, c_i) = \frac{1.7a_j\left(1 - P_{ij}\right)\left(P_{ij} - c_j\right)}{1 - c_j}$$

here $P_{ij} = P_{ij}\left(\theta_i; a_j, b_j, c_j\right)$. To apply the Newton-Raphson method an initial ability estimate, which is referred as θ^- is chosen. And then, the new value of the ability level is estimated by using

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

θ^+ will be closer to the ability parameter than θ^-. The new value is redefined as θ^- and these steps are repeated multiple times until θ^+ and θ^- are equal at a specified level of precision (Kolen & Brennan, 2014).

Multidimensional IRT equating. Using UIRT procedures with multidimensional data is likely to increase the error. The same is true for equating. To prevent this problem, MIRT equating methods have been developed. These methods are quite new.

Two observed score and one true score procedure have been developed by Brossman and Lee (2013). The first observed score procedure is a direct extension of UO equating method, and is called as “Full multidimensional IRT Observed score equating method”. The second observed score procedure is an approximation of UO equating. This method equates multidimensional exams by using unidimensional IRT equating principles and referred as “Unidimensional approximation of multidimensional IRT observed score equating method”. The last method is a true score equating method and is referred as “Unidimensional approximation of multidimensional IRT true score equating method”. This method is an extension of UT equating method.

Under simple structure MIRT model (SS-MIRT), both observed score (Lee & Brossman, 2012) and true score methods (Kim, Lee & Kolen, 2020) have been developed. And, multidimensional equating methods under the bi-factor models have been proposed for both observed score (Lee & Lee, 2016) and true score equating methods (Lee et al., 2015). Also, for observed score and true score testlet response, model MIRT equating methods have been developed by Tao and Cao (2016). Some comparison studies have been conducted recently in the literature by using multidimensional IRT equating methods (Lee, Lee & Brennan, 2014; Peterson & Lee, 2014; Zhang, 2012). In this study, the next part will mainly focus on SS-MIRT observed score (SMO) equating and the other part will represent a summary of SS-MIRT true score (SMT) equating.

SS-MIRT equating. SS-MIRT framework is different from complex MIRT framework in many aspects. First of all, in SS-MIRT framework each item is associated with only one ability and this makes the calibration process much easier than complex MIRT framework. Second, the correlations between abilities are estimated in SS-MIRT framework but from complex MIRT perspective this is not the case. Third, the complex MIRT framework requires a scale linking procedure even for random groups design, when the SS-MIRT framework does not. Moreover, the SS-MIRT allows explicit interpretation of dimensionality. When items are grouped

according to a pre-specified dimension, such as content domains or item formats, it is considered appropriate to use SS-MIRT framework to equate test forms.

SS-MIRT observed score equating method. Lee and Brossman (2012) developed SS-MIRT observed score (SMO) equating. In their study they used mixed format tests consisting of MC (multiple choice) and FR (free response) items. The three-parameter logistic model was used for MC items when the graded response model was used for FR items. And, the study consisted of two abilities (θ_1, θ_2) which were associated with MC and FR items, respectively.

There are some basic assumptions of SS-MIRT model: (a) each item in the test measures an ability, which corresponds to a specific item type or content domain, and these abilities are correlated, (b) each of the groups of items that are associated with the same ability can be modeled by using unidimensional IRT model.

To conduct a SMO equating, first, items are calibrated using SS-MIRT model. Based on these estimated item parameters, the conditional observed score distributions for each ability ($\theta_1 \text{ and } \theta_2$) is obtained for each form. The conditional observed score distributions of each ability ($\theta_1 \text{ and } \theta_2$) are computed respectively as below:

$$f_1(x_1|\theta_1) = \Pr(X_1 = x_1|\theta_1) \quad \text{and} \quad f_2(x_2|\theta_2) = \Pr(X_2 = x_2|\theta_2)$$

By using a recursive formula (Hanson, 1994; Lord & Wingersky, 1984) the conditional distributions can be found. Total observed score is a sum of weighted scores of different item types or different content domains, $X = w_1X_1 + w_2X_2$. Where X is defined as total observed score. And then, the conditional total score distribution can be computed, under the local independence assumption, as:

$$f(x|\theta_1, \theta_2) = \Pr(X = x|\theta_1, \theta_2) = \sum_{X=w_1X_1+w_2X_2} f_1(x_1|\theta_1) f_2(x_2|\theta_2),$$

where the summation is taken over all possible pairs of $w_1X_1 + w_2X_2$ that gives a particular total score x. To obtain a marginal total score distribution, conditional total score distributions are aggregated over a bivariate ability distribution, $g(\theta_1, \theta_2)$, as below:

$$f(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x|\theta_1, \theta_2) g(\theta_1, \theta_2) d\theta_1 d\theta_2.$$

When a quadrature point set is used for ability distribution, integration is replaced with the summation, as:

$$f(x) = \sum_{\theta_1} \sum_{\theta_2} f(x|\theta_1, \theta_2) q(\theta_1, \theta_2),$$

where $q(\theta_1, \theta_2)$ is the density for a particular θ_1 and θ_2 pair, and $\sum q(\theta_1, \theta_2) = 1$. The marginal observed score distribution is computed for both Form X and Form Y. Traditional EQ equating is then conducted to equate the two forms using the two marginal observed score distributions.

Lee and Brossman (2012) applied SS-MIRT method to a mixed format test that consists of two different item type MC (multiple choice) and FR (free response) items, which are associated with different abilities (θ_1,θ_2). And, there was a correlation between these abilities. According to Lee and Brossman's (2012) results, SMO equating method performed better than the UO equating method. Results from their study represented that as the correlation between abilities (dimensions) decreases, bias in SS-MIRT method tends to decrease, while the opposite happens for UO equating method. That is, when multidimensionality increases, the SS-MIRT method performs better while UIRT method's performance gets worse.

Equating under SMO equating method provides many advantages. First, it gives the opportunity of easy calibration of items due to the capability of using UIRT model. Second, it provides enhanced interpretation of the results unlike the very complex interpretation of many MIRT methods. Moreover, it allows to use score weights effectively (Lee & Brossman, 2012).

SS-MIRT true score equating method. The logic used to perform SMT equating is that under the SS-MIRT framework, only one ability is needed to estimate the probability of a correct response on the item (Kim, Lee & Kolen, 2020). Therefore, equating two forms can be conducted by using a subset of the test that measures the same construct. And then, the final equating relationship can be determined by combining these subset equating relationships. Specifically, the SMT equating method can be conducted by following the steps below:

1. Conducting item calibration for both forms using the SS-MIRT framework;
2. Conducting UT equating for each dimension separately by using the item parameters that are obtained from step 1;

3. Obtaining the weighted equated scores of each dimension to compute the equated composite scores for each combination of section scores. This can be explained as below:

$$eq(x|x_1, x_2, \dots, x_m) = w_1 eq_1(x_1) + w_2 eq_2(x_2) + \cdots + w_m eq_m(x_m)$$

where w_m is the weight of equated score for dimension m and, $eq_m(x_m)$ is the equated equivalent of x_m score for dimension m.

4. For the new form, estimating a multivariate observed score distribution for the m dimension scores,

5. Summing weighted equated composite scores to obtain the final equated score. This can be computed as below:

$$eq(x) = \sum_{X=w_1X_1+w_2X_2+\cdots+w_mX_m} f(x_1, x_2, \dots, x_m|x)\big(w_1 eq_1(x_1) + w_2 eq_2(x_2) + \cdots + w_m eq_m(x_m)\big)$$

where $f(x_1, x_2, \dots, x_m|x)$ is defined as the relative frequency (Kim, 2018; Kim, Lee & Kolen, 2020).

Equipercentile Methods

Equipercentile (EQ) equating methods have been developed for common item nonequivalent groups design. These methods focus on total score and common item score distributions, instead of focusing on the means, standard deviations and covariances that are the focus of many other methods. Hence, a synthetic population is usually required to conduct EQ equating with common item nonequivalent groups design (Kolen & Brennan, 2014). Frequency estimation and chained equipercentile (CE) equating methods are some of the EQ equating methods. In this study, one of the investigated equating methods is CE equating method, which is described in detail below. Additionally, in this study the abbreviations of EQ and CE are preferred in order to be compatible with current studies on multidimensional equating.

Chained equipercentile equating. In this method, first scores on Form X are converted to the scores on common items by using population 1. And then, the sores on common items are converted to the scores on Form Y by using population

2. As the last step, these two conversions are chained together and hence, a conversion of the scores of Form X to the scores of Form Y is obtained (Kolen & Brennan, 2014). Marco et al. (1983) referred to this method as the “direct equipercentile” method. On the other hand, Livingston et al. (1990) referred to this method as the “chained equipercentile” method. As the reason, they pointed out that this method consists of two separate equipercentile equatings and these are linked by the anchor test (common items). This method is entirely different from many other methods according to the way of using the common item scores. For example, it does not use the new and old form distributions. Instead, it equates the new form to the common items in new form sample and the old form to the common items in the old form sample. Also, this method assumes that the equating relationship between each form and common items is the same across the populations (Livingston et al., 1990). To explain this method specifically, the steps are given by Kolen and Brennan (2014) in detail below:

1. Use EQ equating method to convert scores of Form X to the scores of common items for population 1. And, refer this equating function as $e_{V1}(x)$.
2. Use again EQ equating method to convert scores of common items to the scores of Form Y for population 2. And, refer this equating function as $e_{Y2}(v)$.
3. To equate Form X scores to Form Y scores, first use $e_{V1}(x)$ to convert a score of Form X to a score of common items and then use $e_{Y2}(v)$ to convert the resulting common item score to a score on Form Y.

Consequently, these steps point out a composed function which involves a chain of two equipercentile conversions:

$$e_{Y(chain)} = e_{Y2}[e_{V1}(x)]$$

It should be noted that, to conduct the EQ equating, the marginal distributions for scores on X and V in Population 1 and the marginal distributions for scores on Y and V in Population 2 are required only (Kolen & Brennan, 2014).

Item Bias and Test Bias

Item bias occurs when examinees, from different groups, have equal ability level, but have different probability of answering an item correctly. The same definition applies to bias in the test. In psychometric literature, for statistical observations, a term other than bias was suggested to be used. And then, differential item functioning (DIF), which is quite apart from judgmental or interpretive meaning and use of bias, came into use. DIF is based on the simple observation that an item has different statistical properties for the examinees, whose abilities are checked to be matched but are from different groups (Holland & Wainer, 1993).

There are two directions of DIF in the literature: unidirectional and bidirectional. In unidirectional DIF condition, all DIF items in the test favor only one subgroup. On the other hand, in bidirectional DIF condition, while some of DIF items favors one subgroup, the other DIF items favor another subgroup. In this dissertation only unidirectional conditions were used.

According to the literature, there are two types of differential item functioning: uniform and non-uniform DIF. If the association between the item response and the group is constant for all ability levels then it is called uniform DIF; otherwise, it is called non-uniform DIF when the mentioned association is a function of the ability levels (Hanson, 1998). On the other hand, differentiation of a, b, and c parameters in groups can also be used to define DIF type.

While DIF focuses on group invariance at item level, differential test functioning (DTF) refers to the sum of DIF in the test. That is, DTF concerns invariance at the raw test score level. Equating invariance, DTF, and DIF deal with invariant relationships across subgroups. Specifically, DIF focuses on invariance at item level, DTF focuses on invariance at raw score level, and equating invariance focuses on invariance at reported score level (Huggins, 2012). To sum up, DTF represents the aggregate effect of all DIF items in a test and can be computed as below:

$$DTF = E_F(T_F - T_R)^2,$$

where E is the expectation which may be taken over the reference (R) or the focal (F) group, T_F is the true score for being a member of the focal group, and T_R is

the true score for being a member of the reference group. The greater the difference between T_F and T_R, the greater the DTF (Oshima, Raju & Flowers, 1997).

A potential for item bias exists, when subgroups of examinees have different underlying abilities. That is, multidimensionality may cause item bias. True and nuisance ability dimensions need to be identified carefully to decide an item as biased in multidimensional framework. The true ability (or abilities) is the ability that the test is designed to measure. There may be more than one true ability intended to be measured by the test. On the other hand, the nuisance ability helps examinee to solve the particular item but it is not intended to be measured by the test. If the distribution of the nuisance ability differs across subgroups, the potential for item bias exists. Item impact and item bias should not be confused by each other. To summarize, when subgroups of examinees differ in their performances on the abilities that are intended to be measured by the test items, item impact occurs. That is, true differences in the ability cause difference in the results. As opposite, in item bias the reason of the difference in the results is the nuisance abilities (Ackerman, 1992).

Relationship between Equating Invariance and DIF

Equating invariance and DIF both use the population invariance to assess the level of group dependence of statistical functions. Equating invariance does not hold when the differential difficulty of the two test forms is not consistent across subgroups. On the other hand, DIF examines whether the item score is invariant across subgroups. When DIF focuses on the item scores, equating invariance focuses on the reported scores. The DIF analysis alone is insufficient to explain the effect of DIF on the reported scores. As two aspects of fairness, equating invariance and DIF should be addressed together (Dorans, 2004).

According to the literature in CTT (Classical Test Theory), differential difficulty across test forms results in equating dependence (Cook & Petersen, 1987; Dorans, 2004). With a similar logic, for IRT, it can be hypothesized that in the absence of DIF in all test items, equating invariance should be ensured (Huggins, 2012). However, this is too general to explain the relationship between DIF and equating invariance. Hence, Huggins (2014) addressed this relationship in IRT perspective.

There are more studies needed to explain the relationship between DIF and equating invariance clearly. MIRT is a rapidly developing study field and it is considered to be important to clarify the relationship between DIF and equating invariance in MIRT perspective. Based on the hypothesis that DIF in items has an effect on equating in MIRT, in current dissertation different equating methods were used to investigate this effect in multidimensional perspective.

Review of Relevant Research

For psychometric literature, MIRT equating is an essential subject area. However, there is a limited number of studies conducted on this subject yet. These studies have been carried out in recent years. Some of these studies are for developing MIRT equating methods, and the remaining studies are for comparison of various equating methods. According to the relevant literature, there is only two study (Kim, Lee & Kolen, 2020; Zhang, 2012) using the CINEG design so far. Also, in the present dissertation, the CINEG design was used.

The number of studies investigating the relationship between DIF and equation is also rare. In particular, there is no study on the relationship between DIF and MIRT equating. The main research topics of the current dissertation are multidimensional IRT equating, the relationship between equating and DIF, and population invariance of equating. The main studies in the literature on these research topics, as well as current studies dealing with similar issues were investigated and then summarized below.

Lee and Brossman (2012) developed SMO equating method based on the expectation that using multidimensional equating methods with multidimensional data would give more accurate results. In this study, they used multidimensional tests that consisted of items, each of which matched with a single proficiency. These proficiencies, which were separate but correlated, were associated with two different item formats, multiple choice and free response. Because of consisting two different item formats, tests were referred as mixed format tests. Both real data and simulation studies were used and according to the results using SMO equating method gave adequate results when the data structure was multidimensional. Also, it should be noted that SMO equating method outperformed the traditional UIRT method in the presence of multidimensional data.

Kim, Lee and Kolen (2020) were developed SMT equating method. In their study, they conducted four studies with different data types: (1) real data, (2) simulated data, (3) pseudo forms data, and (4) intact single form data with identity equating. They also added four different equating methods to compare the performances of these methods with SMT equating method: (a) EQ equating with presmoothing, (b) UT equating, (c) UO equating, and (d) SMO equating. According to the results of SMT equating method behaved similarly to the four equating methods. Moreover, SMT equating method produced more accurate results compared to UIRT methods. SMT equating outperformed UT equating method in three studies consistently. Hence, these results supported the use of multidimensional equating methods with multidimensional data.

Brossman and Lee (2013) created two observed and one true score equating methods for use in multidimensional IRT. The first observed score equating method, which is a direct extension of UO equating, is referred as “Full MIRT observed score equating method”. The second observed score equating and the true score equating methods used unidimensional approximation procedures to conduct equating under unidimensional principle. These procedures are referred as the ‘‘Unidimensional Approximation of MIRT True Score Equating Procedure”, and the ‘‘Unidimensional Approximation of MIRT Observed Score Equating Procedure’’. In addition to these methods UO and UT equating methods, and EQ equating method were used in this study. Here, EQ equating method was used as a criterion because this method does not violate the unidimensionality assumption. According to the results of this study, MIRT equating methods performed more similarly to the equipercentile method than the UIRT methods. This may be caused due to the violation of unidimensionality assumption of UIRT methods.

Lee and Lee (2016) developed bi-factor multidimensional IRT (BF-MIRT) observed score equating method and then evaluated its performance compared to UO equating method. They conducted equating for mixed format tests containing MC and FR items and treated these item formats as different dimensions. According to the results of this study, two methods acted similar in many cases.

Lee and his colleagues (Lee et al., 2015), developed BF-MIRT true score equating method. In this study, eight different equating methods, which contain both true and observed score equating methods, based on dichotomous IRT, polytomous

IRT, testlet response model, and bi-factor model were used. These methods were compared to EQ equating. According to the results, while true and observed score equating methods based on dichotomous and bi-factor model gave similar results to each other, methods based on polytomous model gave similar results to the EQ equating.

Tao and Cao (2016) created true and observed score equating methods for testlet response theory (TRT) model under multidimensional framework. Results of this study indicated that when local item dependence is at moderate or high level, testlet response theory methods produced more accurate equating results compared to the unidimensional methods.

All of the above-mentioned studies are method development studies for MIRT equating. The next ones are comparison studies, which compared these methods to various equating methods and evaluated their performances. Zhang (2012), used three MIRT equating methods (the full MIRT observed score equating, the unidimensional approximation of MIRT true score equating, and the unidimensional approximation of MIRT observed score equating methods) with various linking methods under the CINEG design. According to the results of this study, the unidimensional approximation of MIRT true score equating method performed best across all linking methods and all group distributions conditions.

Lee, Lee and Brennan (2014) compared six equating methods under random groups design: (1) full MIRT observed score equating, (2) unidimensionalized MIRT observed score equating, (3) unidimensionalized MIRT true score equating, (4) unidimensional IRT observed score equating, (5) unidimensional IRT true score equating, (6) equipercentile equating. Four main conditions were investigated in this study: test length, sample size, form difficulty difference, and correlation between dimensions. The FMIRT (full MIRT) observed score equating and identity equating were used for criterion equating relationships. Finally, results were evaluated by investigating bias, standard error, and overall error. According to the results, the full MIRT observed score equating method performed better than other methods especially for the condition of low correlation between dimensions. Another important result of this study was that even for multidimensional tests the UIRT equating methods gave adequate equating results. As a final result, for small form

difference, large sample size, and long test length, equating of multidimensional tests gave more accurate results.

Peterson and Lee (2014), introduced a full MIRT observed score equating method for mixed format tests. The other equating methods used in this study were bi-factor observed score equating, UO and traditional EQ equating methods. Identity and EQ equating were used in this study to obtain criterion equating relationships. In general, for the data sets that showed more dimensionality, the multidimensional methods performed better. On the other hand, for the data sets, which were unidimensional, unidimensional methods performed better.

According to the literature review, there is a limited number of studies about the relationship between DIF and test equating. One of these studies was conducted by Dorans (2004). This study showed that subpopulation is very important for both DIF and equating. However, DIF analysis focuses on items while equating focuses on reported scores. Therefore, DIF analyses are not adequate alone to assess fairness. DIF and equating invariance analyses should be done together.

One of the most basic studies aimed at dealing with the relationship between DIF and equating invariance more clearly and understandably belongs to Huggins (2014). In this study, it was emphasized that DIF and equating invariance analyses are both affected from subpopulation invariance. But in psychometric literature these two issues have been treated separately. From this point of view, in this study, the effect of DIF on equating invariance was investigated. According to the results, differential form DIF have an important effect on equating invariance of test equating.

Atalay Kabasakal and Kelecioglu (2015), investigated equating results of various methods for conditions where DIF existed and did not exist in multilevel item response models and traditional item response models perspective. They assessed the performances of the methods according to equating errors. According to the results of this study, for some conditions the results of multilevel item response models, for the other conditions the results of traditional item response models were better. Thus, it could not be said that any of these methods outperformed the others.

Demirus and Gelbal (2016), compared various methods under conditions that contained and did not contain DIF. They used RMSD index, which was computed

by using the difference between ability estimates, to compare the performances of the methods. Based on the results they pointed out that when DIF existed on common items, mean-mean method produced the biggest equating error, while mean-sigma method produced the smallest error. When DIF did not exist in the common items, mean-sigma method produced the biggest error, while Stocking-Lord (SL) and Haebara methods produced the smallest.

Yurtcu and Guzeller (2018) compared the errors of mean-mean, mean-sigma, Haebara, and SL methods by adding DIF to some specific items. Based on the results of this study, they pointed out that adding DIF to test items caused an increase in the errors.

In the studies mentioned above, the change in the equating results was observed for the conditions where DIF exists. However, to clarify the relationship between DIF and test equating, population invariance must be examined. Unlike these studies, in this dissertation, population invariance is emphasized to explain the relationship between DIF and test equating. There are some basic studies about population invariance in psychometric literature. These are mentioned below.

Dorans and Holland (2000), conducted a study containing the statistics they developed to measure the population invariance of equating. In their opinion, when two tests are not equitable, it is very likely that the linking functions are not invariant across different subpopulations. According to the ideas that they pointed out, there is not any equating function, which is completely population invariant. Instead, the population dependence is small enough to be ignored. In this study, they developed two indices to compute the difference between subpopulations and the overall population in equating dependence. The first index was the root mean square difference (RMSD), which computes the difference for the subpopulation linking function and the overall linking function. The second index was the root expected mean square difference (REMSD), which is computed by averaging the values before taking the square root in RMSD.

Yang (2004), conducted an application study to investigate group invariance of linking functions over subgroups defined by region. Hence, the root expected squared difference ($RESD_j$), which is one of the group-to-overall indices, was developed by this study. In conclusion, linkings across regions were group invariant.

Dorans, Liu and Hammond (2008) conducted an exploratory study to clarify the role of population invariance in equating results. They investigated equating results under small and large ability differences. According to the results, they emphasized that population invariance is a prerequisite for equating. In other words, if there is a lack of equating invariance, it can be said that the linking is not an equating. On the other hand, meeting population invariance cannot guarantee that score interchangeability feature of equating has been achieved.

Huggins and Penfield (2012) conducted a study that presented available methods for evaluating population invariance in linking and equating. In this study, they pointed out that population invariance in linking and equating is very important to ensure the validity and fairness of test scores. This study was an instructional module that provides an overview of relevant researches about the population invariance of linking and equating. Methods for evaluating equating invariance were categorized and introduced under two topics: omnibus or group-to-overall methods, and conditional or unconditional methods. In this study, they also developed an equating invariance method called RSD_j, which is a conditional version of $RESD_j$ (Yang, 2004) method.

Powers and Kolen (2014), investigated equating results of various equating methods by creating group differences. In this study, frequency estimation, chained equipercentile, IRT true score equating, and IRT observed score equating methods were used. According to the results, when group differences increased, equating results became dissimilar among equating methods, and also behaved biased. They pointed out in this study that an equating method should be selected by considering the size of group differences, the likelihood of the violation of the equating assumptions, and the error associated with the equating method.

Dorans, Lin, Wang and Yao (2014), investigated the linking relationships between latent test scores. From this point of view, they tried to present how observed score linkings were affected from these latent linking relationships. The effects of correlation between latent dimensions, and difference in test content on linking functions were examined. In conclusion, the results of this study focused on the relationship between the correlation between latent dimensions and equitability of test scores as test content differentiates.

According to the literature review, the number of studies examining the relationship between DIF and equating is quite rare. And, there is no study to address this relationship in MIRT perspective yet. In this respect, the results of this research are very important for psychometric literature. In addition, the number of studies conducted with multidimensional equating methods is also rare. And, previous studies were focused on equating errors while the current study examines the performance of MIRT, UIRT and EQ equating methods in terms of population invariance. Therefore, the results of the current study will make significant contribution to the MIRT methodology.

Chapter 3
Methodology

In this study, different equating methods were compared under the presence of DIF in multidimensional tests. In current applications, examinees are need to have more than one ability to answer test items correctly. Hence, multidimensionality and multidimensional test equating have become interesting research topics. In the presence of multidimensionality, it is very difficult to prevent DIF. According to literature review, there are not any studies which investigate the relationship between differential item functioning and multidimensional test equating yet. Therefore, it is very important to find out the effects of DIF on test equating in this perspective.

Research objective of this study was comparing MIRT, UIRT and EQ equating methods in the presence of various simulation conditions. These conditions are related to group differences and DIF. Simulation procedures which are conducted to find out the relationship between DIF and equating are explained in detail below.

Data Preparation

In this study, MIRT, UIRT and EQ equating methods were compared based on their population invariance measures obtained in different simulation conditions. To investigate the relationship between test equating and DIF in multidimensional perspective, two-dimensional data structure was used. In order to obtain the SS-MIRT model, both dimensions were created by using 3-PL model. Specifically, item parameters of two dimensions were generated by using the distributions of an item parameter pool and then correlation between two dimensions was formed to obtain the SS-MIRT structure. These generated parameters served here as true item parameters. And, data were generated by using true item parameters. The weights of 1 was used for the scores in both dimensions, and hence total score range obtained as 0-80.

CINEG (Common item nonequivalent groups) design was used, in which two equated samples were assumed to come from different populations with bivariate normally distributed abilities. 1.681 (41x41) pairs of bivariate quadrature points and weights were used. For both abilities, the range of theta values were assumed to be

from -4 to +4 According to Kolen and Brennan (2014), common items should be at least 20% of the total test items. Based on this, common items were formed to be 25% of the total test items. Based on 100 replications, four equating procedures conducted in this study: SMO, UO, UT, and CE.

Composite scores consist of two or more content areas (e.g., a mathematic test is formed of algebra and geometry) or two or more item types (e.g., a test is formed of multiple-choice and free-response items). Content areas mentioned here and item types measure different abilities which are correlated to provide a single score of achievement (Kolen, Wang and Lee, 2012). In the present study, two dimensions were considered to measure two different content areas and it is assumed that there is a correlation between these two dimensions. Therefore, it is suitable to call this model “Simple Structure-MIRT Model”. Composite scores were formed by summing scores from two content areas. For the purpose of this study, only raw score equating was conducted.

In this study, two subgroups were considered: reference group and focal group. For most situations, it is very difficult for sample sizes of subgroups to be approximately equal to each other. Therefore, subgroups were disaggregated to have different percentages. While focal group was formed of 25% of total group, reference group was formed of 75% of total group. For many equating situations, 3.000 can be argued to be a sufficiently large sample size (Lee et al., 2012). In this study, sample size was fixed to 1.000 for focal group, 3.000 for reference group, and hence total group sample size was fixed to 4.000.

In the current study, DIF items were only generated in the common items. According to Huggins’ (2012) study, the effect of frequency of DIF items on equating dependence was small in IRT equating. Based on this result and by also considering the purpose of the current study, frequency of DIF items were fixed to a particular value. And, this amount was decided to be 20%. Even though it may be unlikely for a test to include 20% DIF common items after examination of item fairness (Lee & Zhang, 2017), it was anticipated to be useful to investigate this amount to represent a severe scenario. In addition to this, moderate level of DIF items were used to investigate the effect of DIF on test equating. DIF in b parameters is generally seen as the primary concern of problematic DIF in operational testing situations (Huggins, 2012). Thus, all DIF was simulated in b parameters.

Simulation Factors

Ability effect size (ES). According to Lee et al. (2012), ability differences between groups have the potential impact on equating results with CINEG design. When the effect size (ES) is .05 or .1, all methods have acceptable equating results. Thus, .1 was considered as small group difference for the current study. According to Kolen and Brennan (2014), around .3 or more standard deviation unit of mean group differences can cause large differences on equating results with CINEG design. And hence, in the present study, .3 was considered as large group difference. To sum up, in this study, two levels of ability effect size measures were considered: .1 for small group differences and .3 for large group differences. Kolen and Brennan (2014) also stated that the difference in group standard deviations can cause differences among equating methods at least as great as the difference caused by mean group differences. And hence, standard deviations were fixed to 1 across all study conditions to prevent any other differences in equating methods. Means, standard deviations, and correlations for the combined groups were computed by using the formulas which were expressed in detail in Dunlap's (1937) study. To obtain the ability effect size measures for population 1 and 2, distributions of focal and reference groups were formed as below:

- *ES* = .1:

$(\theta_1, \theta_2)_{Old}$ ~ *BN*(0, 0, 1, 1, *ρ*) and $(\theta_1, \theta_2)_{New}$ ~ *BN*(.1, .1, 1, 1, *ρ*)

$(\theta_1, \theta_2)_{Old_F}$ ~ *BN*(-.3, -.3, .9, .9, *ρ*) and $(\theta_1, \theta_2)_{New_F}$ ~ *BN*(-.2, -.2, .9, .9, *ρ*)

$(\theta_1, \theta_2)_{Old_R}$ ~ *BN*(.1, .1, .99, .99, *ρ*) and $(\theta_1, \theta_2)_{New_R}$ ~ *BN*(.2, .2, .99, .99, *ρ*)

- *ES* = .3:

$(\theta_1, \theta_2)_{Old}$ ~ *BN*(0, 0, 1, 1, *ρ*) and $(\theta_1, \theta_2)_{New}$ ~ *BN*(.3, .3, 1, 1, *ρ*)

$(\theta_1, \theta_2)_{Old_F}$ ~ *BN*(-.3, -.3, .9, .9, *ρ*) and $(\theta_1, \theta_2)_{New_F}$ ~ *BN*(0, 0, .9, .9, *ρ*)

$(\theta_1, \theta_2)_{Old_R}$ ~ *BN*(.1, .1, .99, .99, *ρ*) and $(\theta_1, \theta_2)_{New_R}$ ~ *BN*(.4, .4, .99, .99, *ρ*)

Correlation between dimensions. Lee and Brossman (2012) investigated SS-MIRT and UIRT procedures in their study and indicated that when the correlation was .8 or higher both methods might produce adequate results. Lee and Lee (2016) used MIRT and UIRT equating methods in their study and they concluded that at

the correlation level of .8 or higher, UIRT and MIRT equating results might be similar. According to this literature review, .8 as correlation between two dimensions was decided to be used as a benchmark for this study. As a higher level, .95 was decided to be used in current dissertation to represent approximately unidimensional situations. .5 or lower level of correlation may be rarely seen in actual situations. But in some testing conditions distinct abilities may be included to the exam to be tested (e.g., language exams). To investigate this type of conditions, .5 were used. To sum up, tree levels of correlation between dimensions were used: .5, .8 and .95. It should be emphasized that correlation levels were checked after data was generated.

Differential form DIF. In CTT (Classical Test Theory) equating, equating dependence occurs when the differential difficulty of the two tests changes across two groups (Dorans, 2004). According to Huggins' (2012) results, in IRT equating, identical DIF in anchor items across forms does not have an impact on equating invariance but when DIF in anchor items differs across forms, equating dependence occurs. In current study, this understanding was discussed in MIRT equating perspective. Based on this, two conditions were formed in this current study: DIF that is identical across test forms and DIF that is differential across test forms. In the first condition, DIF was equivalent across two test forms. That is, for an anchor DIF item, the magnitude and the direction of DIF were the same at two forms. In the second condition, while true DIF were not added to the items in the old form, true DIF were added to the items in the new form. In current study CINEG design was used. According to this design, non-equivalent groups were formed across test forms. The second sub condition were simulated to reflect the possibility of DIF that might be differential across test forms in the non-equivalent groups. A third condition was that DIF did not exist in both forms. Thus, in this research, the results of the equating methods were compared in terms of equating invariances for conditions where DIF did not exist (no-DIF), DIF existed in both forms (DIF in both forms) and DIF existed only in one form (DIF in new form only).

Magnitude of DIF. To investigate the effect of DIF on test equating, moderate DIF level was aimed to be used in current dissertation. Using items which include large amount of DIF may not be practical because these items are mostly detected well by specialists and generally omitted from the tests. Hence, the level of DIF magnitude was used in the current study is .6 and this magnitude was chosen based

on Educational Testing Service's (ETS) classification (Dorans & Holland, 1992). SS-MIRT allows each dimension to be modeled using UIRT. And, this gives us the opportunity of using UIRT while adding DIF to the items. Hence, in both dimensions, uniform DIF was formed by increasing b parameters as 0.6 unit to create moderate level of DIF. It should be emphasized that DIF magnitude was checked after data was generated.

Direction of DIF. The aggregated effect of small or moderate DIF items on test equating may be significant. This understanding can be extended to the MIRT equating. Unidirectional DIF items were used in this study to express the combination of DIF effects. Hence, all DIF items in the test favored to R group in both dimensions.

Data Generation

Distributions of *a*, *b*, and *c* parameters of an item parameter pool were used in this study. The item parameter pool was obtained from multiple forms of a language exam of a large-scale testing program applied in the United States. The item parameter pool was consisted of 800 *a*, *b*, and *c* parameters (based on 3-PL model). The distribution of these item parameters was investigated in detail. The distributions of *a*, *b*, and *c* parameters were shown in Figure 6.

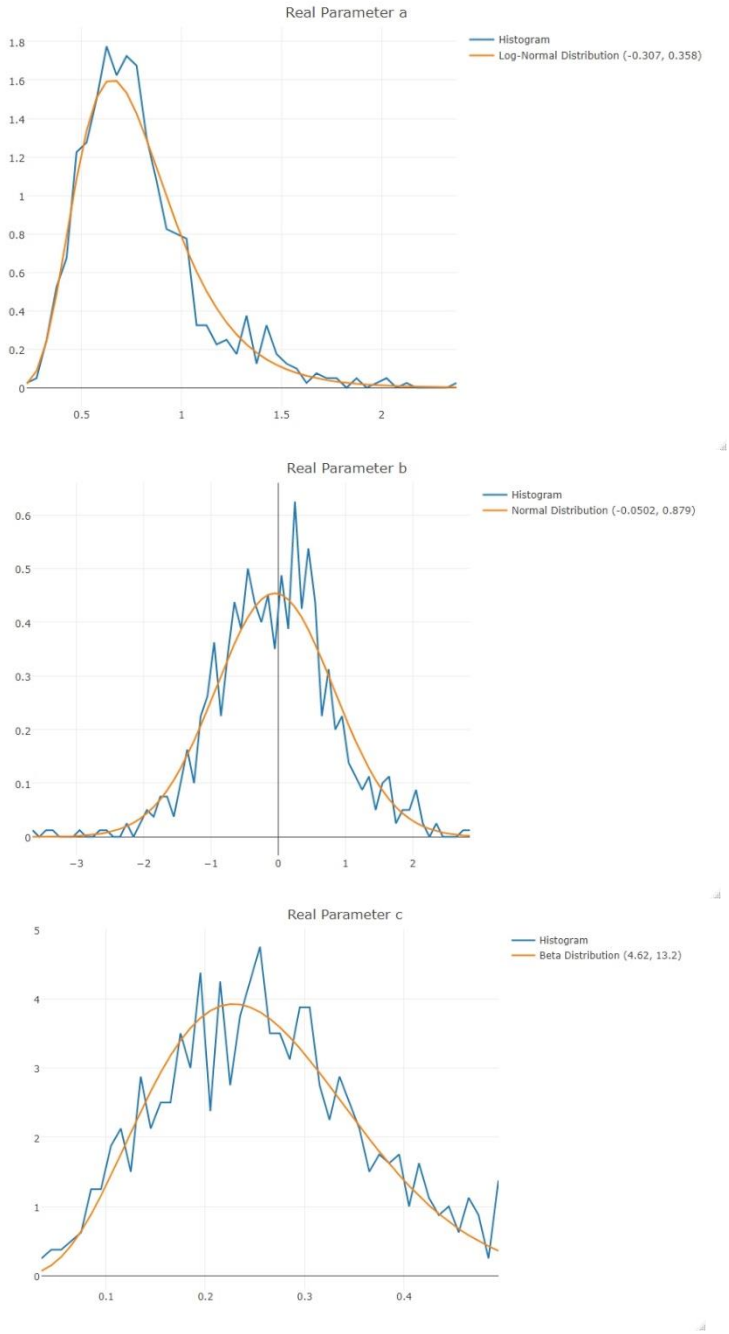

Figure 6. Distributions of real item parameters.

As can be seen in Figure 6, the distributions of *a*, *b*, and *c* parameters nicely fit to log-normal (-0.307, 0.358), normal (-0.0502, 0.879), and beta distributions (4.62, 13.2), respectively. The mean of this beta distribution can be computed as below:

$$\mu = \frac{\alpha}{\alpha + \beta} = \frac{4.62}{4.62 + 13.2} = 0.259$$

The computed mean is appropriate for a 4-choice item test:

After determining the distributions of *a, b,* and *c* parameters, item parameters (*a, b,* and *c*) of this study were generated by using these distributions to represent the characteristics of the item parameter pool. That is, item parameters (item parameters for 20 common items, 60 non-common items for Form 0, and 60 non-common items for Form 1) were generated by using log-normal distribution for *a* parameter, normal distribution for *b* parameter, and beta distribution for *c* parameter. For both common and non-common item parameters, same distributional properties were used to ensure that the common items were a “mini version” of the total test form (Kolen & Brennan, 2014).

Generated item parameters for 40 items (10 common and 30 non-common) represented the first dimension, while the other 40 items represented the second dimension. Correlation was formed between two dimensions to create an SS-MIRT structure for the test form. After generating the item parameters for the old form, similar process was repeated for the new form by considering the common items as the same in both forms. The generated item parameters of two forms were used as the true item parameters for this study. Additionally, DIF were added to the first two items of both dimensions for DIF conditions. Specifically, these items were items 1 and 2 for the first dimension, items 41 and 42 for the second dimension. For items 1, 2, 41, and 42, difficulty parameters of focal group were increased by 0.6 unit to create uniform DIF. All these processes were repeated for 100 times for each simulation condition.

After generating item parameters, ability parameters of both forms were generated by using bivariate normal distributions $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{\theta_1\theta_2})$ which were mentioned in ability effect size conditions. According to these conditions, ES=0.1 and ES=0.3 were used to form group ability differences between two forms for both dimensions. Thus, in accordance with the CINEG design, difference was made between the ability means of the groups that took different test forms. While creating ability parameters, previously specified distributions were used. Specifically, focal and reference groups distributions were determined to form a specified total group distribution for both dimensions. To achieve this, the following equations that were included in Dunlap's study were used in both dimensions.

$$M_T = \frac{rM_r + fM_f}{r + f}$$

Where, M_T is the ability mean of the total group, M_r is the ability mean of the reference group, M_f is the ability mean of the focal group, *r* is the sample size of the reference group, and *f* is the sample size of the focal group. In addition to this, the following equation was used for the calculation of variances:

$$\sigma_T^2 = \frac{r(\sigma_r^2 + \delta_r^2) + f\left(\sigma_f^2 + \delta_f^2\right)}{r + f}$$

where, σ^2 represents the variance, and $\delta_r = M_r - M_T$ represents the difference between the subgroup mean and the total group mean. Finally, the following equation was used for combining correlation coefficients:

$$r_{xy} = \frac{r\sigma_{x_r}\sigma_{y_r}r_{x_ry_r} + r\delta_r\Delta_r + f\sigma_{x_f}\sigma_{y_f}r_{x_fy_f} + f\delta_f\Delta_f}{\sqrt{r\left(\sigma_{x_r}^2 + \delta_r^2\right) + f\left(\sigma_{x_f}^2 + \delta_f^2\right)}\sqrt{r\left(\sigma_{y_r}^2 + \Delta_r^2\right) + f\left(\sigma_{y_f}^2 + \Delta_f^2\right)}}$$

where, *r* and *f* refer again groups to be combined. For any *x* or *y* values in group *m*,

$$x = x_m + \delta_m, \qquad y = y_m + \Delta_m,$$

For any *x* or *y* values in group *n,*

$$x = x_n + \delta_n, \qquad y = y_n + \Delta_n,$$

and,

$$\delta_m = M_{x_m} - M_{x,} \qquad \Delta_m = M_{y_m} - M_{y,}$$

$$\delta_n = M_{x_n} - M_{x,} \qquad \Delta_n = M_{y_n} - M_{y}.$$

As a result, by using these equations, bivariate normal (BN) distributions for focal, reference, and total groups were formed with respect to ES=0.1, and ES=0.3 conditions as follows:

- *Effect Size* = .1:
    - For correlation value of 0.5:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .5) $(\theta_1, \theta_2)_{New}$ ~*BN*(.1, .1, 1, 1, .5)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .5) $(\theta_1, \theta_2)_{New_F}$~*BN*(-.2, -.2, .9, .9, .5)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .48) $(\theta_1, \theta_2)_{New_R}$~*BN*(.2, .2, .993, .993, .48)

    - For correlation value of 0.8:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .8) $(\theta_1, \theta_2)_{New}$ ~*BN*(.1, .1, 1, 1, .8)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .8) $(\theta_1, \theta_2)_{New_F}$~*BN*(-.2, -.2, .9, .9, .8)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .792) $(\theta_1, \theta_2)_{New_R}$~*BN*(.2, .2, .993, .993, .792)

    - For correlation value of 0.95:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .95) $(\theta_1, \theta_2)_{New}$ ~*BN*(.1, .1, 1, 1, .95)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .95) $(\theta_1, \theta_2)_{New_F}$~*BN*(-.2, -.2, .9, .9, .95)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .948) $(\theta_1, \theta_2)_{New_R}$~*BN*(.2, .2, .993, .993, .948)

- *Effect Size* = .3:
    - For correlation value of 0.5:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .5) $(\theta_1, \theta_2)_{New}$ ~*BN*(.3, .3, 1, 1, .5)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .5) $(\theta_1, \theta_2)_{New_F}$~*BN*(0, 0, .9, .9, .5)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .48) $(\theta_1, \theta_2)_{New_R}$~*BN*(.4, .4, .993, .993, .48)

- For correlation value of 0.8:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .8) $(\theta_1, \theta_2)_{New}$ ~*BN*(.3, .3, 1, 1, .8)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .8) $(\theta_1, \theta_2)_{New_F}$~*BN*(0, 0, .9, .9, .8)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .792) $(\theta_1, \theta_2)_{New_R}$~*BN*(.4, .4, .993, .993, .792)

- For correlation value of 0.95:

$(\theta_1, \theta_2)_{Old}$ ~*BN*(0, 0, 1, 1, .95) $(\theta_1, \theta_2)_{New}$ ~*BN*(.3, .3, 1, 1, .95)

$(\theta_1, \theta_2)_{Old_F}$~*BN*(-.3, -.3, .9, .9, .95) $(\theta_1, \theta_2)_{New_F}$~*BN*(0, 0, .9, .9, .95)

$(\theta_1, \theta_2)_{Old_R}$~*BN*(.1, .1, .993, .993, .948) $(\theta_1, \theta_2)_{New_R}$~*BN*(.4, .4, .993, .993, .948)

With respect to the specified distributions, 1000 ability parameters were generated for the focal group, and 3000 ability parameters were generated for the reference group in accordance with the sample size condition for the old form. And, this process was repeated for the new form by creating group ability differences (for ES=0.1, and ES=0.3) as pointed out above. As a final step, all processes were repeated 100 times for both forms.

Finally, item responses were formed by using the generated item and ability parameters with respect to the SS-MIRT model. For each condition, item responses were generated for both forms 100 times. All these steps were performed in the computer program R (R Core Team, 2016) by using the codes which were written by the researcher.

Data Analysis

In this study, the generated item parameters were used as the true parameters and these true item parameters were used to obtain criterion equating relationships which are discussed in the next section in detail. On the other hand, item responses were used in the process of conducting equatings. First, to conduct SMO equating, item parameters were estimated under the SS-MIRT framework using flexMIRT (Cai, 2017). Concurrent calibration was used in this process. In

concurrent calibration, scale linking was carried out at the time of item calibration, hence additional scale transformation was not needed. After the item parameters were estimated on the same scale, SMO equating was conducted three times to equate two forms: first for the focal group, second for the reference group, and third for the total group. These steps were repeated 100 times. For the sake of clarity, the steps of SMO equating are given below.

1. Item parameters of two forms were estimated on the same scale by using concurrent calibration based on SS-MIRT model.
2. Conditional observed score distributions for each dimension were obtained for each form.
3. Conditional total score distributions were obtained for each form using the conditional observed score distributions.
4. A bivariate normal ability distribution was constructed for each form using the corresponding mean, variance, and correlation estimates obtained from flexMIRT concurrent calibration.
5. Marginal observed score distributions were computed for each form by aggregating conditional total score distributions over the bivariate normal theta distribution.
6. Finally, traditional EQ equating was conducted for the two forms.
7. Steps 1 to 6 were carried out for the focal, reference, and total groups.
8. Steps 1 to 7 were repeated 100 times (for all item response files generated).

After the SMO equating was completed for the focal, reference, and total groups, UIRT and EQ equating were conducted using the estimated item parameters which were obtained according to UIRT procedures. Item parameters were estimated under the UIRT framework using concurrent calibration in flexMIRT. After the item parameters were estimated on the same scale, UO, UT and EQ equating methods were conducted for the focal, reference, and total groups. These steps were repeated for each condition 100 times (for all item response files generated). In many equating studies, 100 was used as the iteration number (Kim, Lee & Kolen, 2020; Lee & Brossman, 2012).

After equating results of all methods (SMO, UO, UT, and EQ) were obtained, RMSD and RSD (for both focal, and reference groups) values were calculated using the equating results of focal, reference, and total groups for each condition and each iteration. And then, means of RMSD and RSD values of each method were obtained for each condition by taking the averages over 100 iterations. Finally, RMSD and RSD means of all methods were compared both among each other and with the results of criterion equating relationship. Detailed information on how these indices are calculated is repeated below.

As an additional information, it should be emphasized that CE equating method was used as the EQ equating method in this study. Because, this research was carried out with CINEG design, and among EQ equating methods the most appropriate one for CINEG design was the CE equating method. However, as can be seen from the figures in the findings section, there are fluctuations on RMSD and RSD plots of the CE equating method. This is because this method uses frequency distributions based on number correct scores. To improve the stability of the results, a univariate log-linear presmoothing method, which is called as log-linear pre-smoothed CE equating method, with polynomial degree of 6 was used. All procedures applied to other methods were repeated for this method, and the obtained results were interpreted all together.

Criterion Equating Relationships

When conducting comparison studies with various equating methods, evaluating the results of these methods with each other is not sufficient. For these studies, it is essential to rely on a criterion relationship known to be correct. Unfortunately, a perfect, complete, and objective criterion does not exist in the literature. Therefore, when equating studies are carried out, the equating relationship, that is considered to give the most accurate results depending on the conditions of the study, is selected as criteria. Some of the criterion equating relationships used in MIRT equating literature are EQ, SMO, identity equating, Full-MIRT observed score, TRM-MIRT (Testlet Response Model-MIRT) observed score equating methods.

The structure of the data (MIRT model), the design used (CINEG), the subgroups created (reference, and focal), the methods compared (SMO, UO, UI,

EQ) were important factors taken into account when choosing criterion equating relationship of this study. Considering these factors, the criterion equating relationships were created based on SMO equating method with true item parameters (not estimates). Specifically, SMO equating was performed for focal, reference, and total groups using 100 previously generated true parameters for each condition. Then, 100 RMSD and RSD (RSDF and RSDR) distributions were calculated for each condition based on the equating results obtained. And, RMSD and RSD means were obtained by averaging these 100 RMSD and RSD distributions for each condition. Thus, these RMSD and RSD means for each condition expressed the criterion equalization relationship. It should be emphasized that true parameters were regenerated for each condition and iteration, as Huggins did in her study (Huggins, 2012). Thus, the results were aimed to be generalized to a particular distribution, not to a particular test. This made the research more generalizable.

Using SMO equating with true parameters minimized the measurement and the equating error. Hence, we could compute the ideal equating invariance indices, and monitored ideally how DIF conditions effect equating invariance, and then compared methods according to this ideal effect. At last, we could show practitioners which method was affected most or least or near to ideal. And warned practitioners to be careful about fairness when equating test scores in some specific DIF conditions.

Evaluation Criteria

Selection of evaluation criteria is another important point in evaluating the performance of equating methods. The focus of the study was comparing the methods according to their population invariances. In accordance with the aim of this study, unstandardized RMSD and RSD indices were chosen as evaluation criteria. These indices were calculated after equating processes were conducted for each condition and for each trial. Hence, there were 100 $RMSD(x)$ and 100 $RSD_j(x)$ (100 RSD_R, and 100 RSD_F) distributions of each method for each condition. The averages of these distributions were computed. Finally, RMSD and RSD means of all methods were compared both among each other and with the criterion equating relationship results. Here, mean distributions represented the

estimate of population invariance of each method. To sum up, for each condition, methods were compared based on their population invariance results.

It should be emphasized that the reason for using the $RSD_j(x)$ index in addition to the $RMSD(x)$ index was to prevent the differential form DIF from affecting the smallest group (Huggins, 2012). Both indices used for evaluation criteria explained in detail as below:

$$RMSD(x) = \frac{\sqrt{\sum_j w_j \left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{YP}}$$

The above equation is computed at each x value. Where, j represents different subgroups, P represents the overall group, $e_{P_j}(x)$ represents separate linking functions for specified subgroups, $e_P(x)$ represents the overall linking function, w_j represents the weighting of each subgroup (a proportional representation of each subgroup in the overall group), σ_{YP} represents the standard deviation of Y scores in P (Dorans & Holland, 2000). This index was adapted to CINEG design (unstandardized version) by removing the dominator component. On the other hand, $\mathrm{RSD}_j(x)$ index can be explained as below:

$$RSD_j(x) = \frac{|d_j(x)|}{\sigma_Q}$$

where σ_Q represents the standard deviation of scores in population Q, and $d_j(x)$ represents the difference between a linked score y based on subgroup j's linking function and a linked score y based on the overall linking function at score level x (Huggins & Penfield, 2012). This index gives the standardized distance between the one subgroup's equating function and the overall equating function at one score level (Huggins, 2012). This index was also adapted to this study (to CINEG design) by removing the dominator component. In other words, the unstandardized version of this index was used for this research.

To assess the magnitude of equating dependence, two difference that matters (DTM) criteria were chosen: 0.5 and 1. While 0.5 is a value that has been used in most studies, as stated in Huggins' (Huggins, 2012) study, to better reveal the problematic level of equating dependence also 1 was used. Equating invariance results compared graphically and visually across all conditions by considering both

DTM criteria. Equating methods were evaluated as the best and worst according to the amount of differentiation from the criterion equating relationship, and according to the distance from both DTM criteria. In particular, the equating method, which gave the closest results to the criterion equating relationship results, were defined as the best method. Besides, DTM values were also taken into consideration while interpreting the results.

Chapter 4
Findings

The results of this study are presented in four sections as follows. In the first section, equating methods (multidimensional, unidimensional, and equipercentile equating) are compared with respect to the findings representing the effect of differential form DIF on equating invariance of the methods. In the second section, comparison of equating methods is discussed according to the effect of correlation between dimensions on equating invariance. In the third section, the findings of the effect of group mean ability differences between two forms on the relationship between DIF and equating invariance of the methods are presented and all methods are compared according to these findings. In the final section, equating methods are compared with respect to the findings obtained in all sections.

Research Question 1

What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of differential form DIF on population invariance?

In this section all equating methods are compared with respect to differential form DIF which include three conditions: no-DIF, DIF in both forms, and DIF in new form only. In no-DIF condition, items were generated to have no true DIF. In DIF in both forms condition, DIF was added to the same common items (1, 2, 41, 42) in both forms in the same direction (favored to R group) and in the same amount (moderate level). In DIF in new form only condition, DIF was added to common items in the new form only.

First, RMSD means of the methods for these conditions with 0.5 correlation between dimensions and 0.1 group mean ability difference between two forms (ES=0.1 for focal, reference, and total groups) are shown in Figure 7. Specifically, Figure 7 represents no-DIF, DIF in both forms, and DIF in new form only results, respectively.

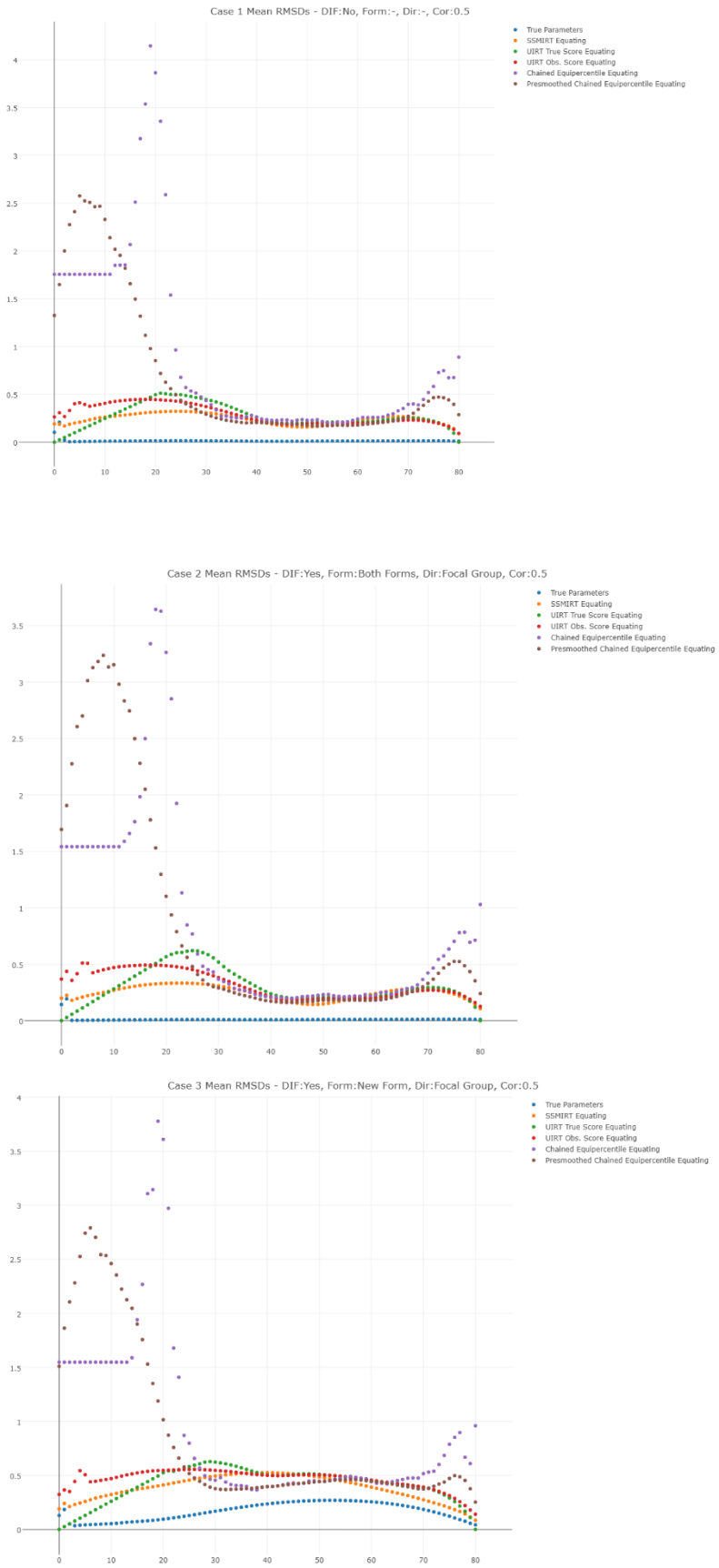

Figure 7. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

In order to see the results more clearly, Figure 8 containing only (M)IRT methods results and Figure 9 containing the zoomed versions of the plots of all methods results were added to the study as below.

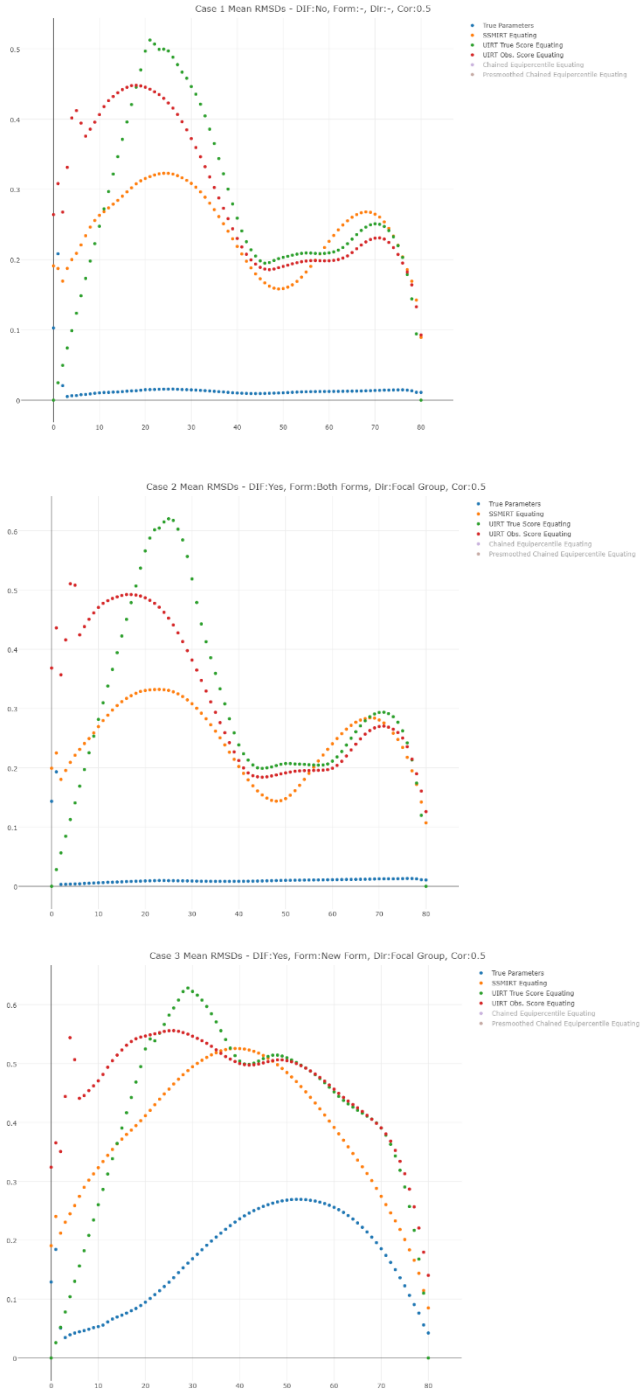

Figure 8. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

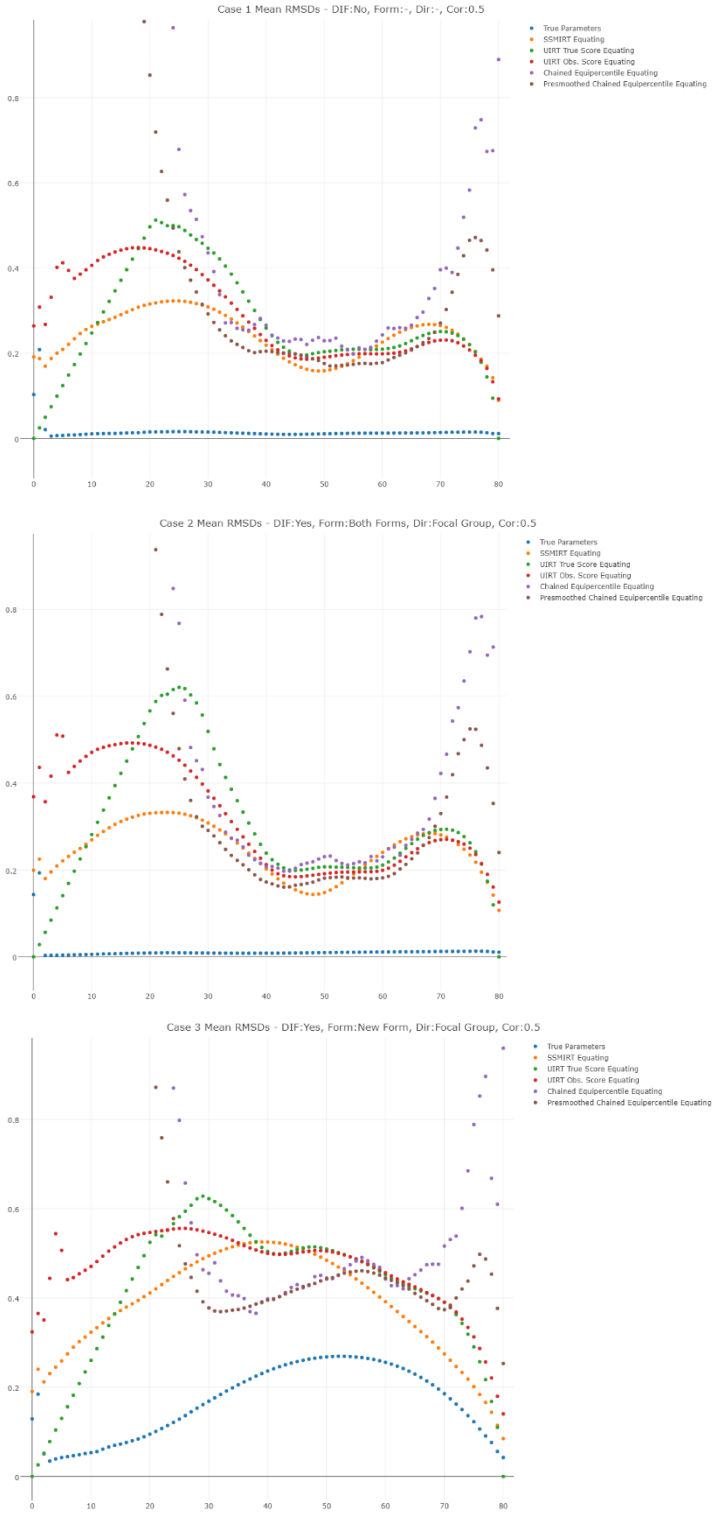

Figure 9. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

As can be seen in Figures 7, 8, and 9, RMSD results are similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and mostly smaller than the DTM of 0.5. However, for DIF in new form only condition, there is an increase in RMSD values for all methods. The results of all methods except the chained and the log-linear pre-smoothed chained methods are around the DTM of 0.5. The results of chained and log-linear pre-smoothed chained methods are smaller for a specific score range. However, these methods produce quite high RMSD values in the score ranges with low frequencies because they conducted equating with respect to the frequency distributions based on number correct scores. Besides, as expected, the log-linear pre-smoothed CE equating method gives smoother results than the CE equating method. Hence, in the next sections of this study, the log-linear pre-smoothed CE equating method results are interpreted. To sum up, according to the results shown in Figures 7, 8 and 9, even if log-linear pre-smoothed CE equating method gives very close results to the criterion equating relationship for a specific score range with high frequencies, for the whole score range SMO equating method gives the closest results to the criterion equating relationship in terms of the distribution and the values.

Figure 10 represents the RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation between dimensions and ES=0.3. Besides, Figure 11 represents only (M)IRT methods results, while Figure 12 represents the zoomed versions of the plots containing all methods results.

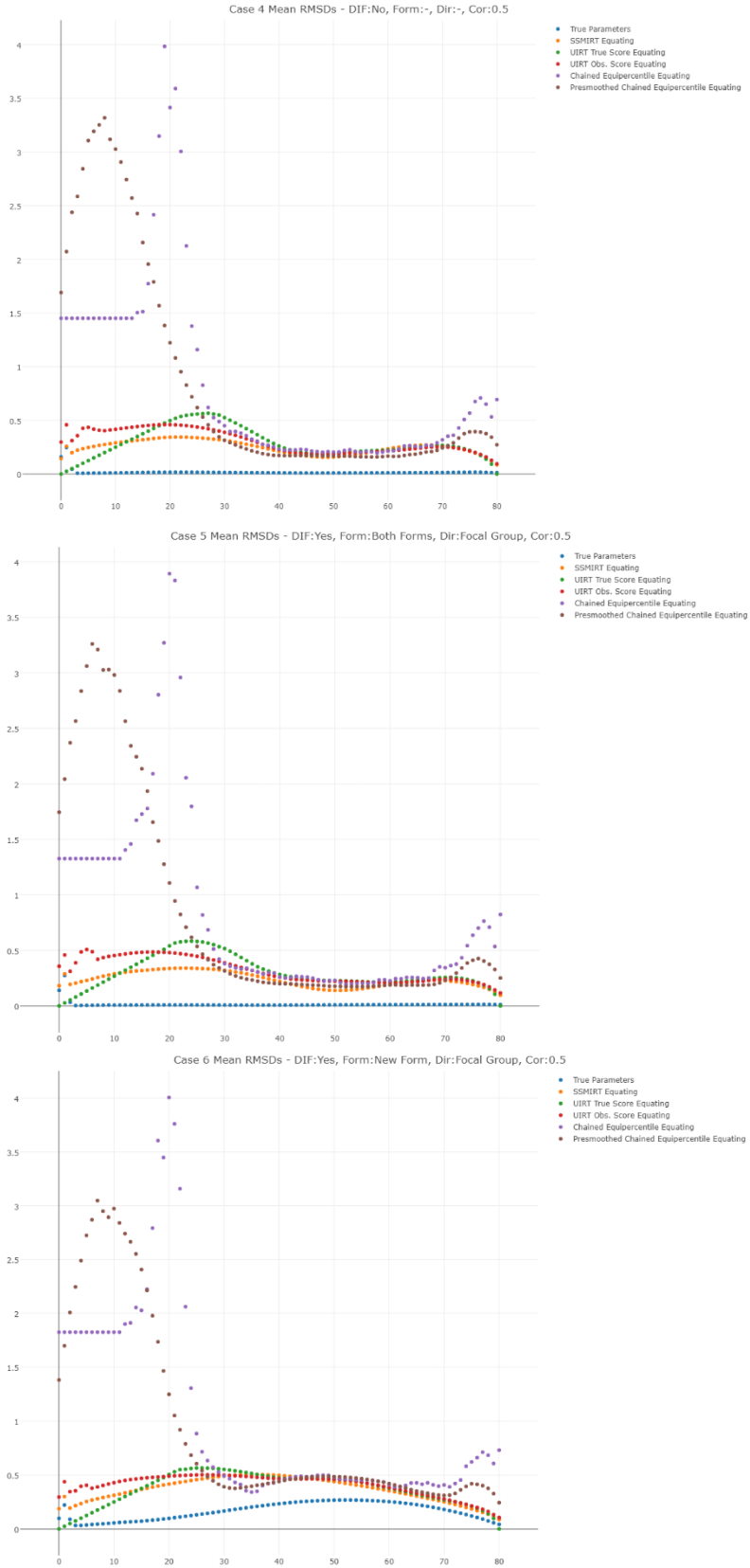

Figure 10. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.3.

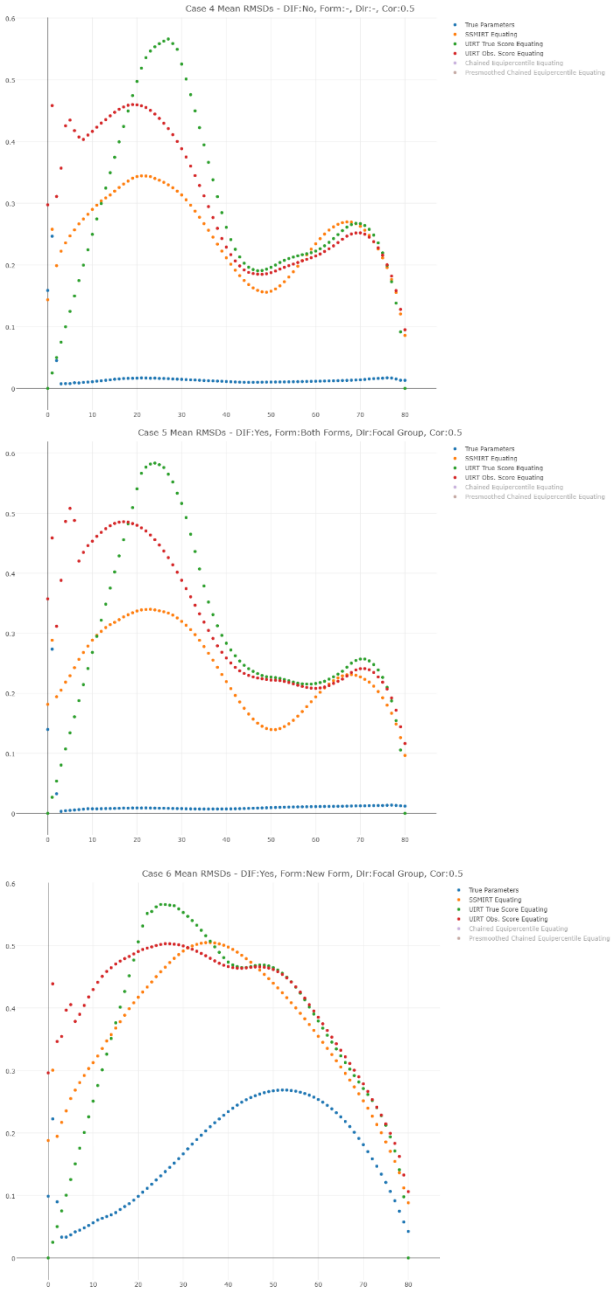

Figure 11. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.3.

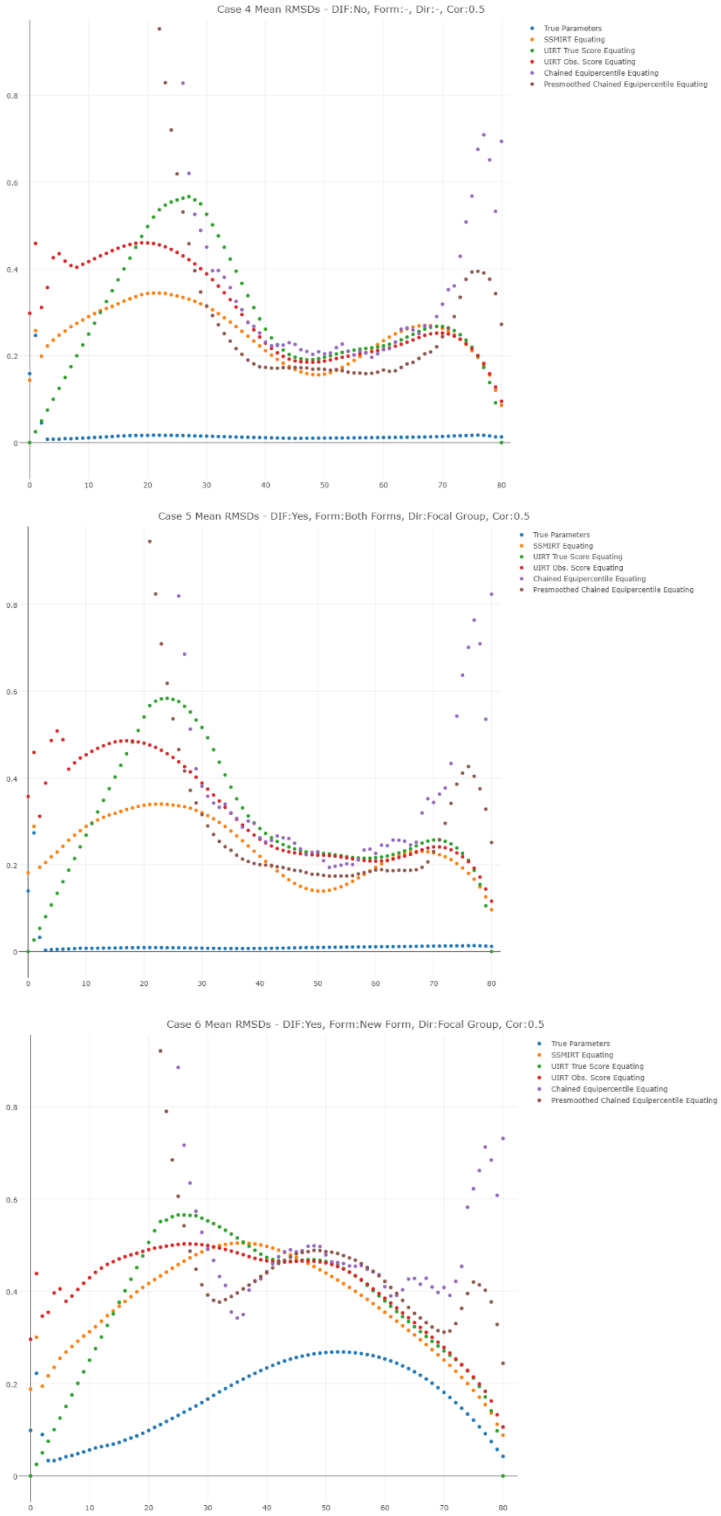

Figure 12. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.3.

According to Figures 10, 11, and 12, the results are quite similar to the conditions where ES = 0.1. Specifically, RMSD results are similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and often smaller than the DTM of 0.5. However, for DIF in new form only condition, there is an increase in RMSD values for all methods. The results of all methods except the CE equating method are around the DTM of 0.5. The results of the CE equating method are smaller for a specific score range however, this method gives quite high RMSD values in the score ranges with low frequencies. To sum up, according to the results shown in Figures 10, 11 and 12, even if the CE equating method gives very close results to the criterion equating relationship for a specific score range with high frequencies, for the whole score range the SMO equating method gives the closest results to the criterion equating relationship in terms of the distribution and the values. Based on these results mentioned for ES=0.3, it can be said that difference in ES has not an impact on the distribution and the values of RMSD results. Similar to the results of ES=0.1, it is seen that the RMSD values of the methods in no-DIF and DIF in both forms conditions are often smaller than the DTM of 0.5, but for DIF in only new form condition, the values are close to the DTM of 0.5 and above. To sum up, in the conditions mentioned, the equating method that gives the closest results to the criterion equating relationship results in terms of the distribution and the values is the SMO.

RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation, and ES=0.1 are represented in Figures 13, 14, and 15 as below. While Figure 13 shows the results of all methods, Figure 14 shows the results of (M)IRT methods only, and Figure 15 includes zoomed versions of the results for all methods.

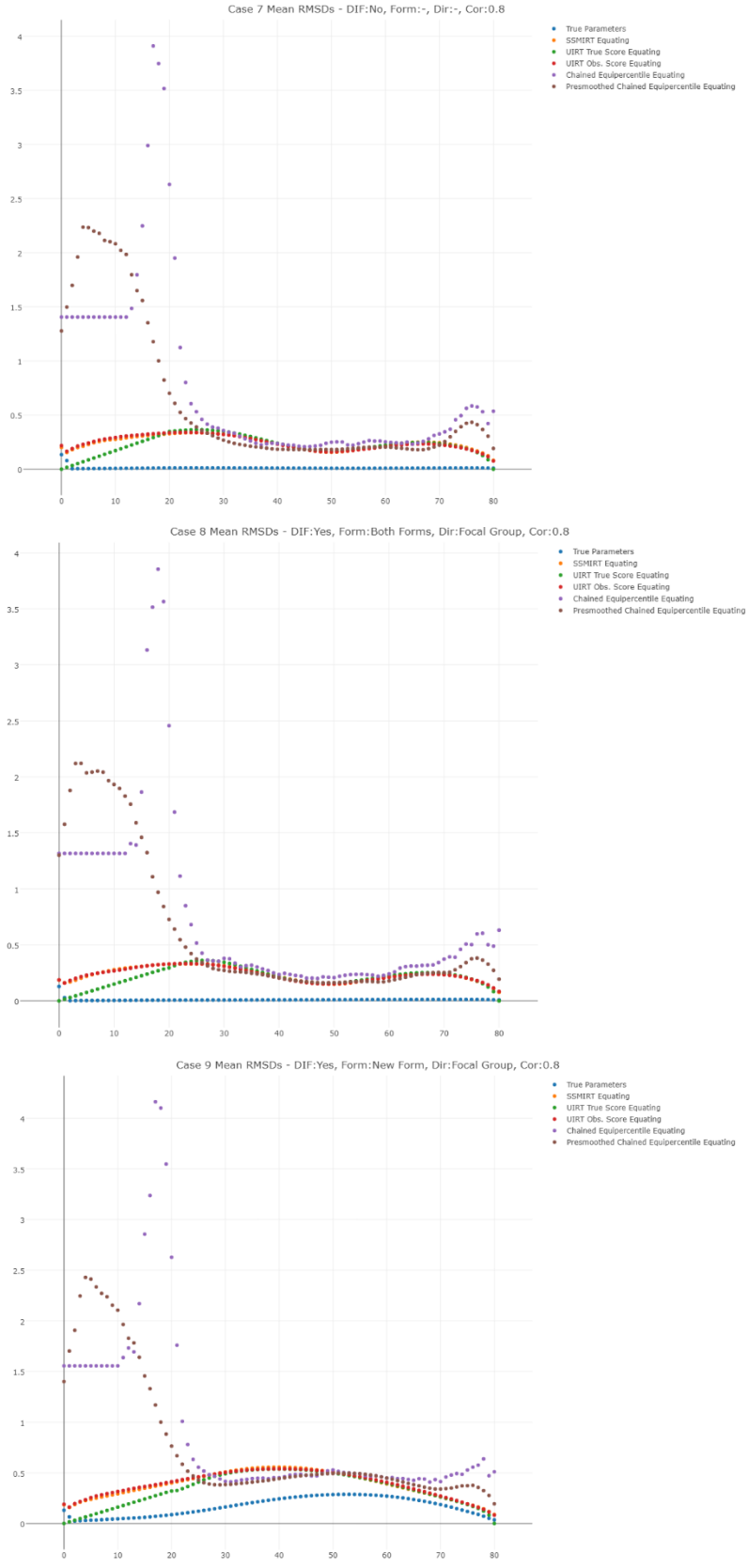

Figure 13. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.1.

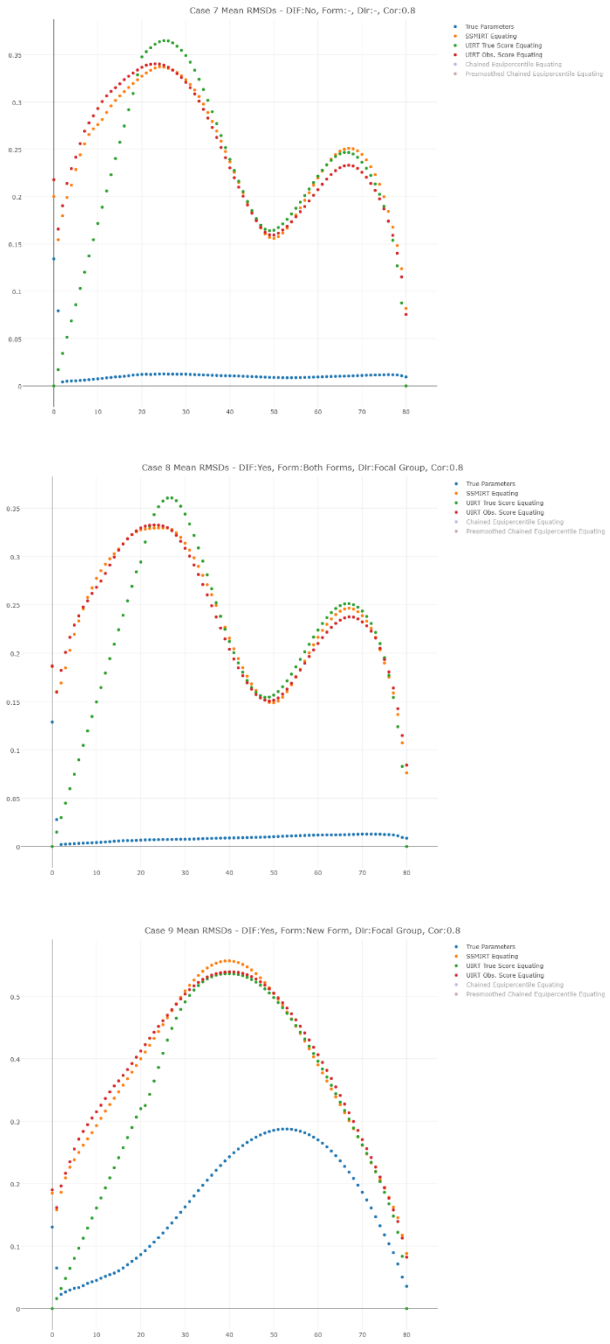

Figure 14. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.1.

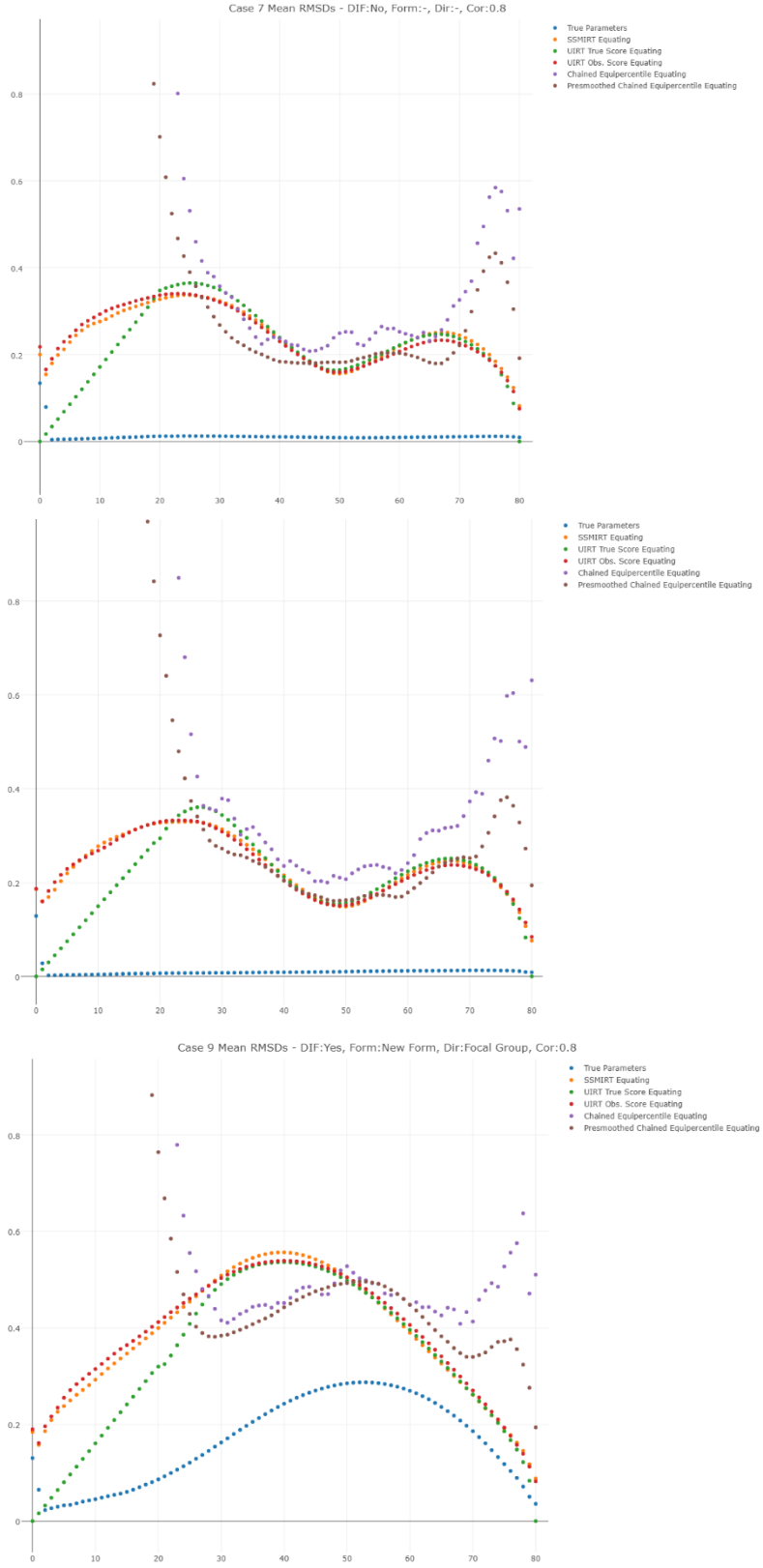

Figure 15. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.1.

According to Figures 13, 14, 15, with 0.8 correlation between dimensions and ES=0.1, the comparison of all equating methods according to no-DIF, DIF in both forms and DIF in new form only conditions is as follows. For no-DIF and DIF in both forms conditions, the results of MIRT and UIRT methods, and the results of the CE equating method for scores with high frequencies are below the DTM of 0.5. When DIF is added to new form only, the RMSD results increase and reach around the DTM of 0.5, and even some of them exceed this value. That is, DIF in one form only causes an increase in the RMSD values of all methods. The results obtained from the conditions mentioned are quite similar in terms of MIRT and UIRT equating methods. In some score ranges SMO, in some others UT, and in some others UO method gives closer results to the criteria. To sum up, it seems difficult to distinguish these methods in terms of the distributions and values of the RMSD results. On the other hand, CE equating method's results are very close to the criterion equating relationship results for the scores with high frequency. However, for the scores with low frequency the RMSD results are quite high.

Figures 16, 17, and 18, which are given below, represent the RMSD means of the methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3. Specifically, Figure 16 includes the results of all methods, while Figure 17 shows (M)IRT methods results only, and Figure 18 shows the zoomed versions of the results of all methods.

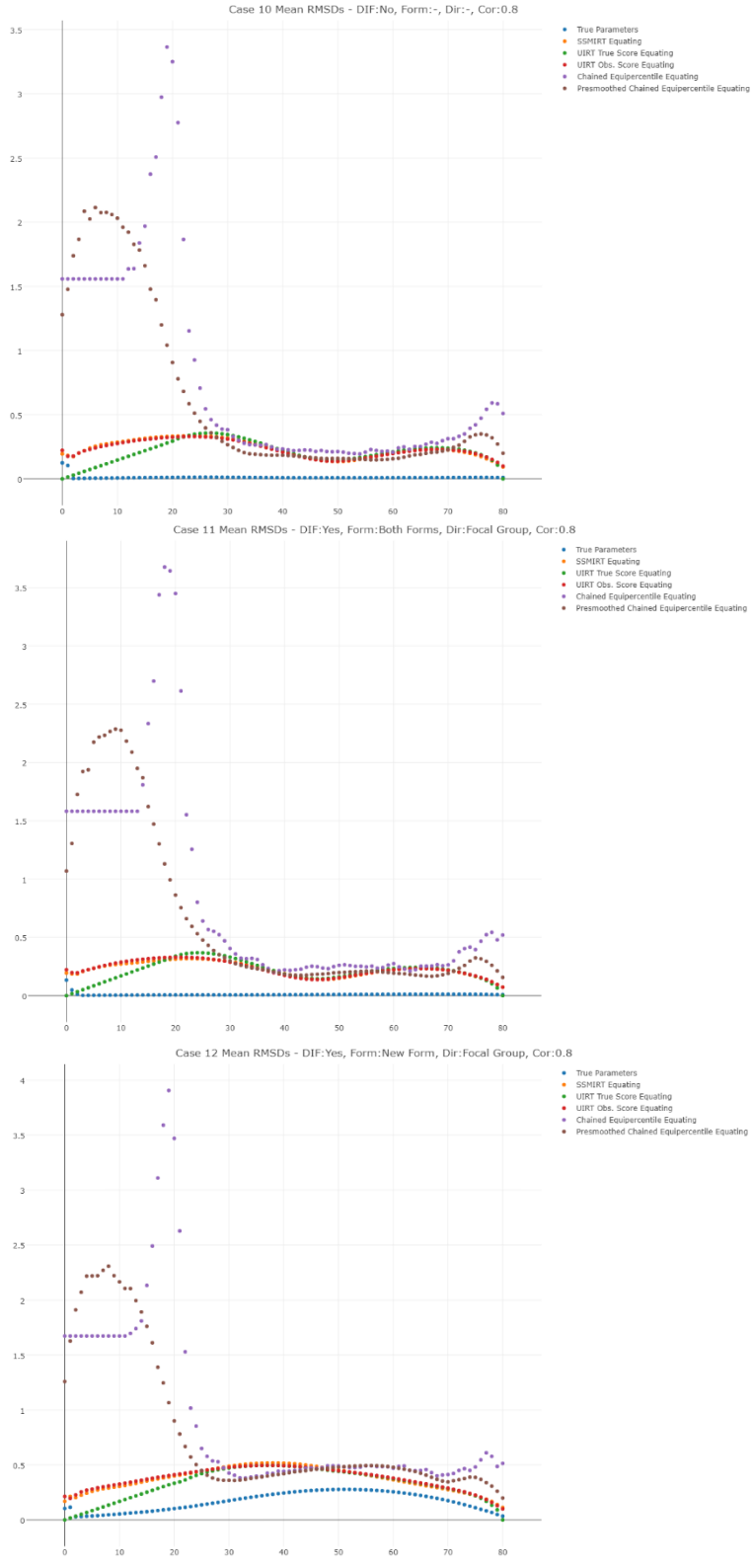

Figure 16. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3.

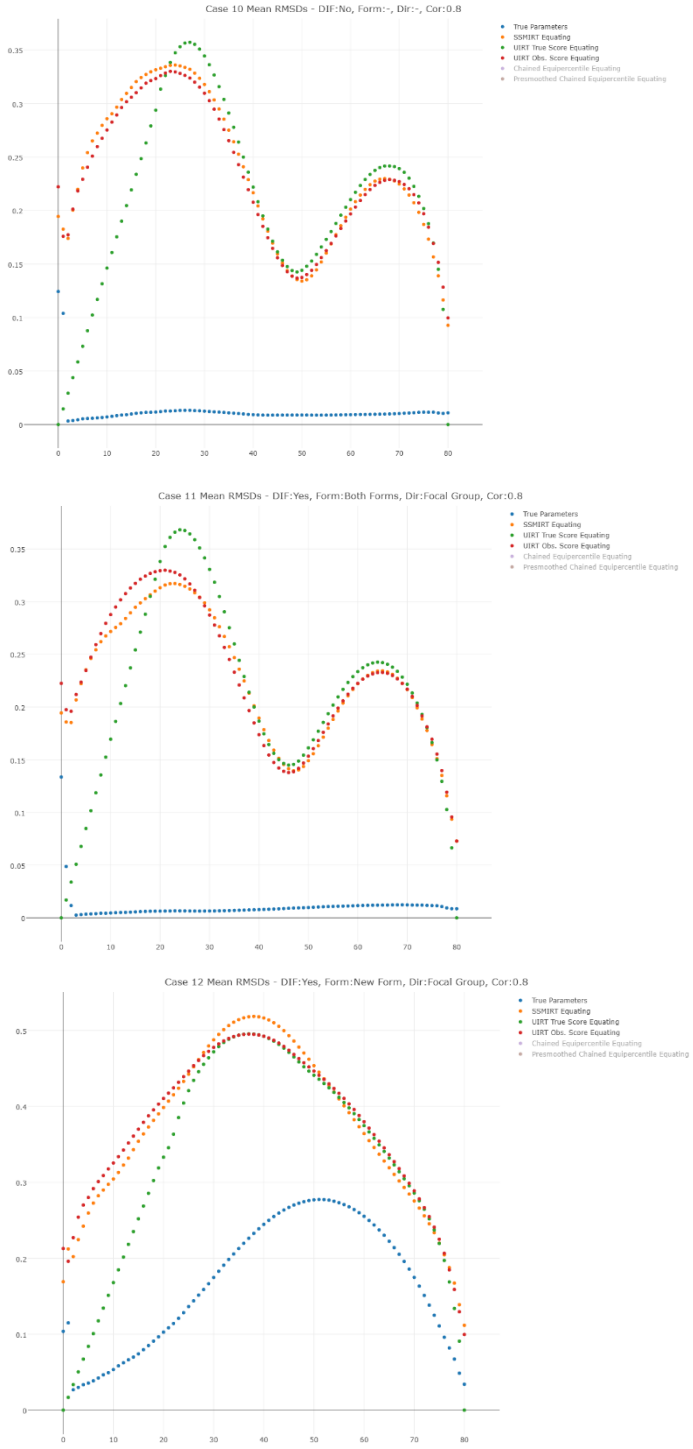

Figure 17. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3.

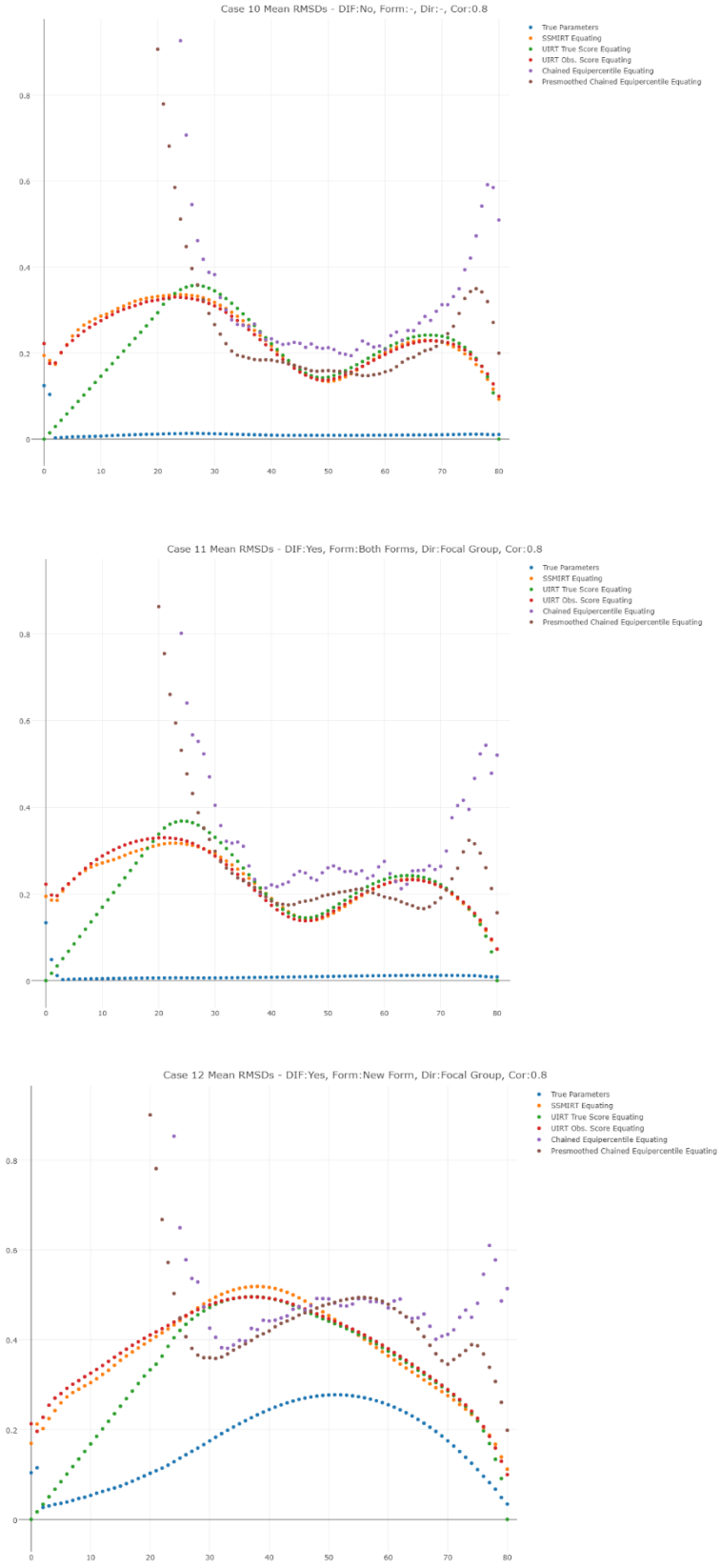

Figure 18. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3.

As can be seen in the Figures 16, 17, and 18, the RMSD results of the methods are similar to those for conditions with ES=0.1. Specifically, for no-DIF and DIF in both forms conditions, the results of MIRT and UIRT methods, and the results of the EQ method for scores with high frequencies are below the DTM of 0.5. When DIF is added to new form only, the RMSD results reach around the DTM of 0.5, and even some of them exceed this value. That is, the RMSD values of all methods increase when DIF added to new form only. MIRT and UIRT equating methods results are quite similar in terms of the values and the distributions. In some score ranges SMO, in some others UT, and in some others UO method gives closer results to the criteria. To sum up, it seems difficult to distinguish these methods in 0.8 correlation with ES=0.3. On the other hand, EQ method results are very close to the criterion equating relationship results for the scores with high frequencies. However, for the scores with low frequencies the RMSD results of this method are quite high.

RMSD means are represented for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation, and ES=0.1 in Figure 19. While in Figure 20 IRT methods are compared with each other, in Figure 21 the zoomed results of all methods are presented. Figures 19, 20, and 21 are given below.

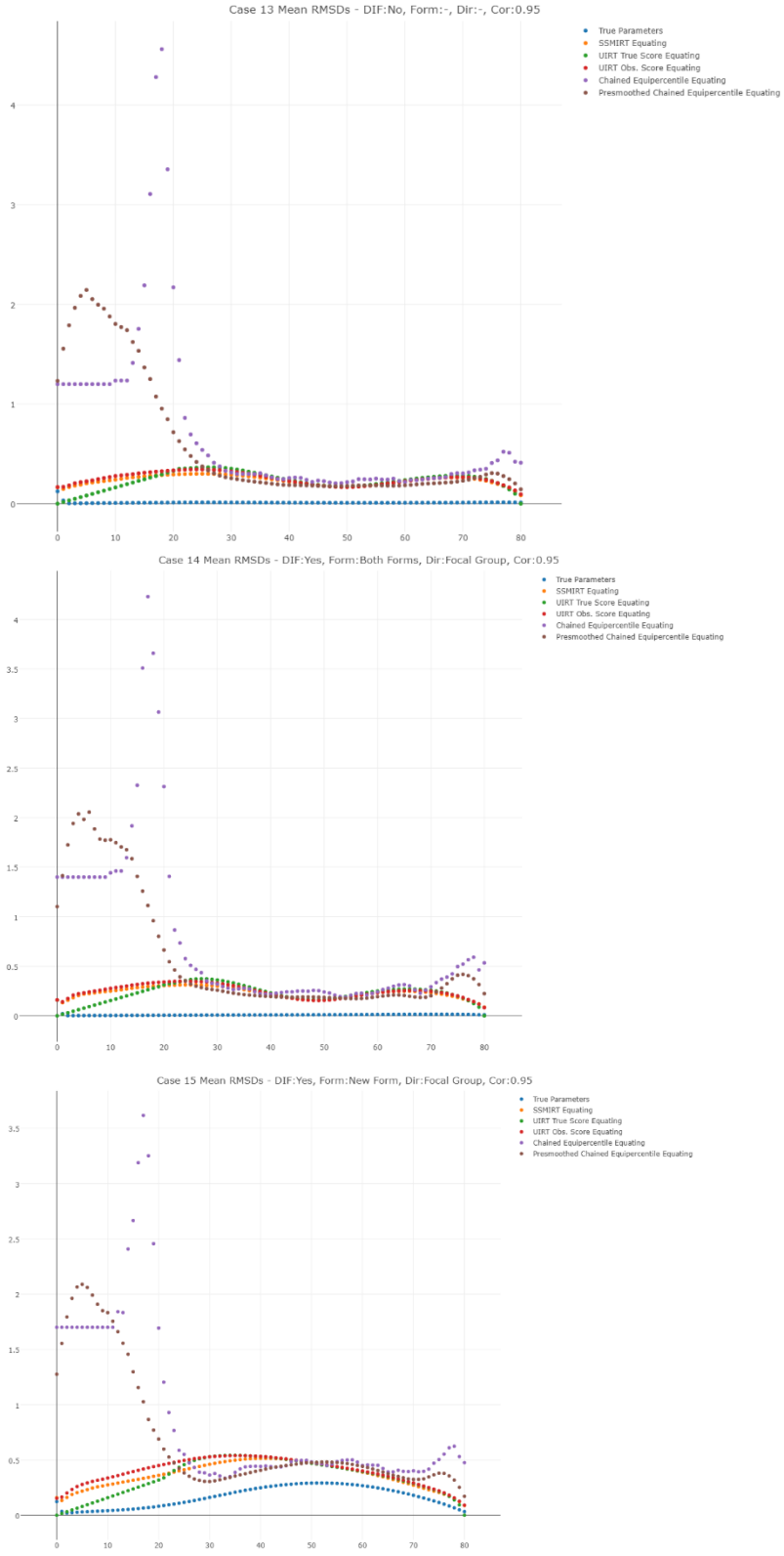

Figure 19. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.1.

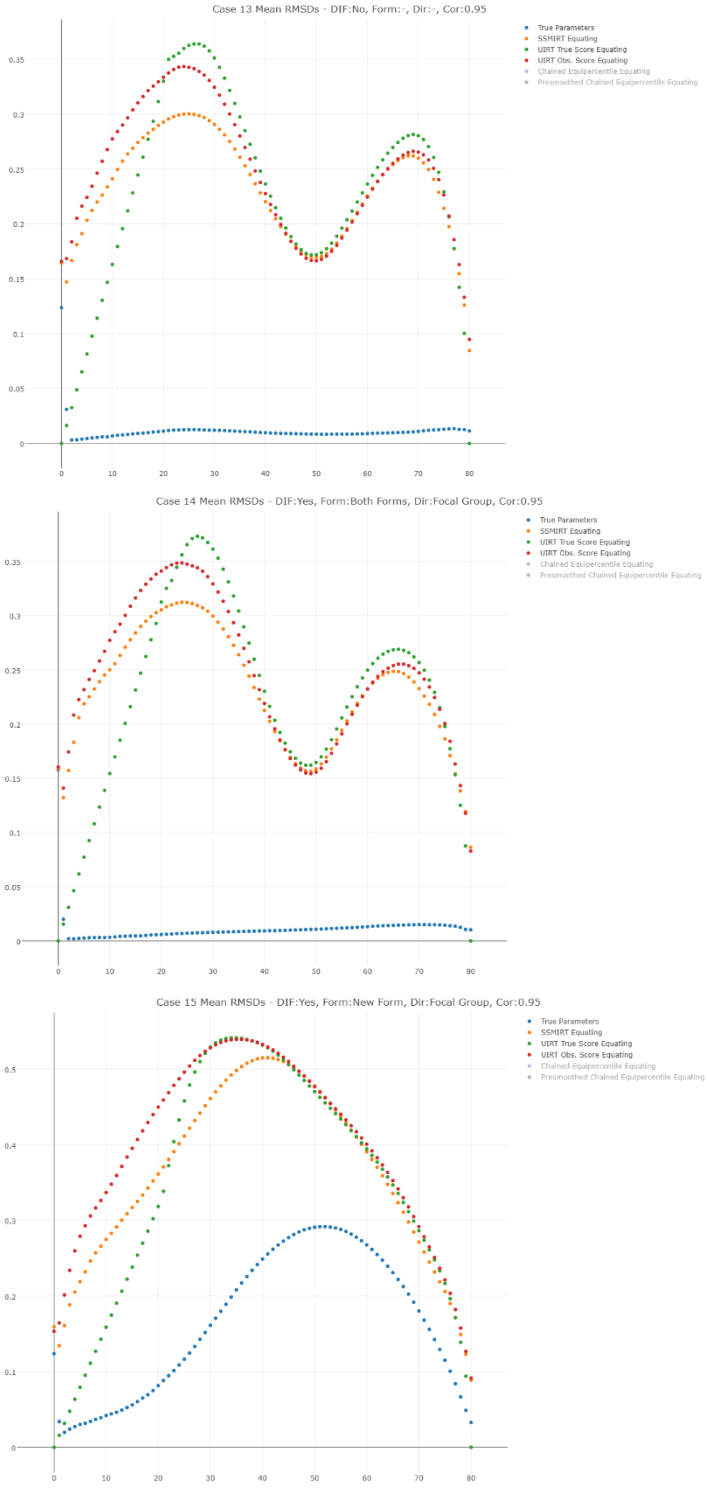

Figure 20. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.1.

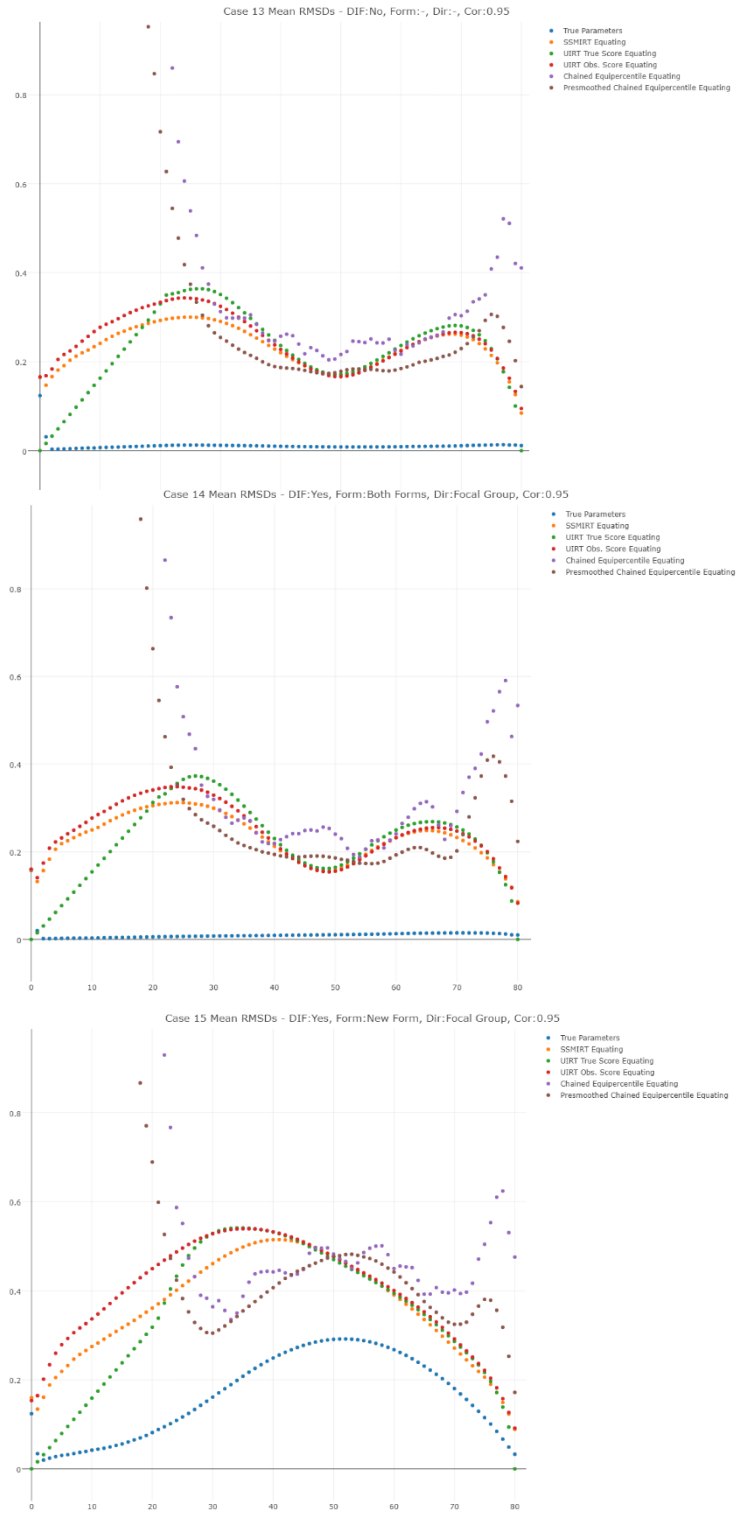

Figure 21. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.1.

Findings with respect to Figures 19, 20, and 21 are as follows. With 0.95 correlation between dimensions and ES=0.1, for no-DIF and DIF in both forms conditions, the RMSD results of MIRT and UIRT methods and results of the EQ method for scores high frequencies are smaller than the DTM of 0.5. When DIF is added to new form only, the RMSD results increase and reach around the DTM of 0.5 and even some results exceed this value. For the cases mentioned, the results are quite similar in terms of MIRT and UIRT methods and the EQ method for the scores with high frequencies. On the other hand, when ES changes to 0.3, results obtained are given in figures below.

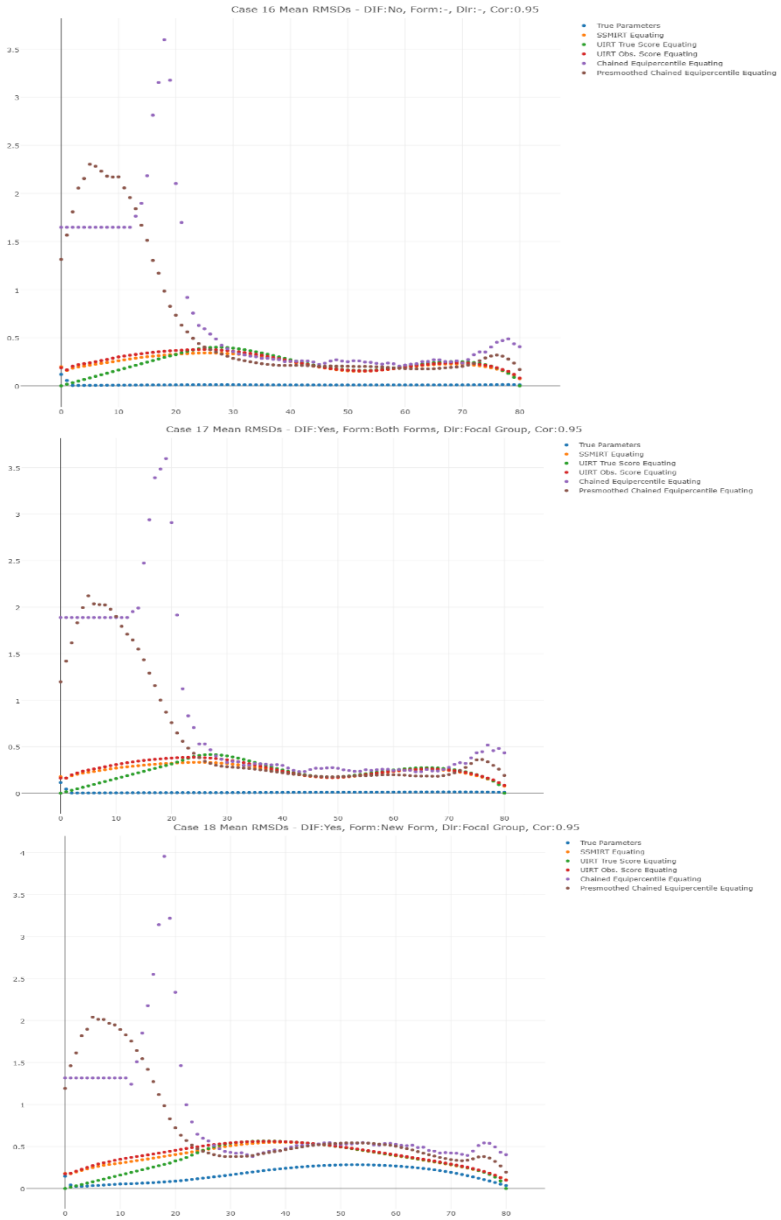

Figure 22. RMSD means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.3.

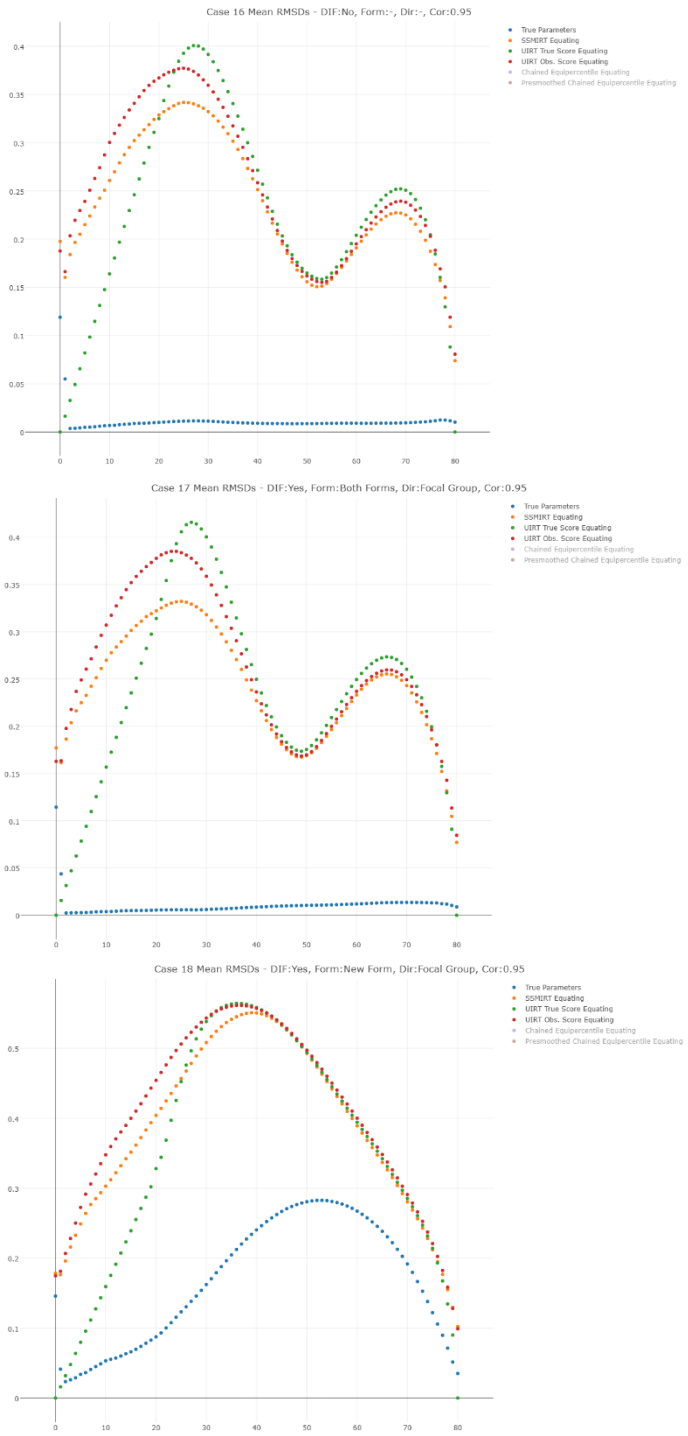

Figure 23. RMSD means of (M)IRT methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.3.

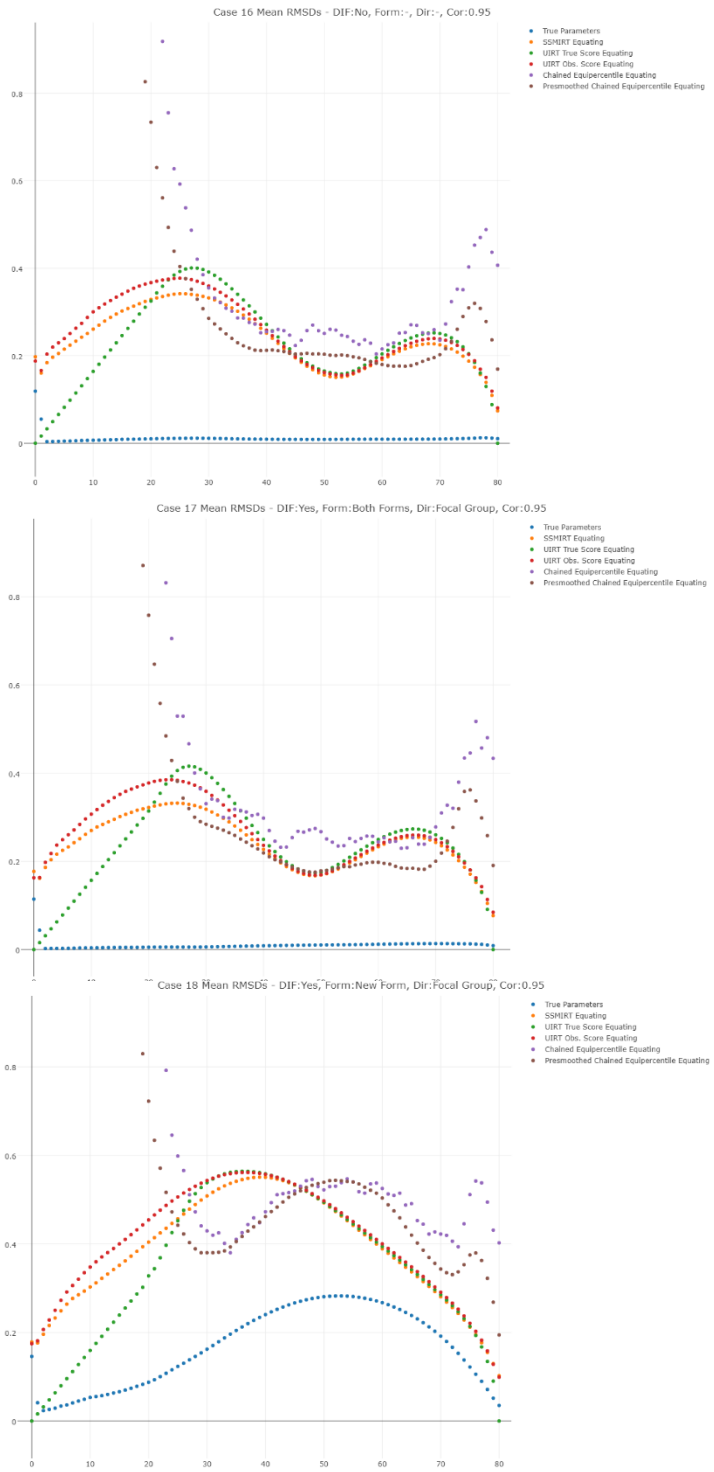

Figure 24. Zoomed RMSD means of all methods for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.3.

According to findings shown in Figures 22, 23, 24, results obtained are similar to those for conditions with ES=0.1. Specifically, with 0.95 correlation between dimensions and ES=0.3, for no-DIF and DIF in both forms conditions, the RMSD results of MIRT and UIRT methods and results of the EQ method for scores with high frequencies are smaller than the DTM of 0.5. When DIF is added to new form only, the RMSD results increase and reach around the DTM of 0.5 and even some results exceed this value. Consequently, MIRT and UIRT results, and the EQ method results for scores with high frequencies are quite similar to each other as in the 0.8 correlation condition.

The other evaluation criterion of this study is RSD, which is one of the group-to-overall conditional equating invariance indices. In the following figures (Figures 25 and 26) RSD_F (for the focal group) and RSD_R (for the reference group) means are represented for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

As can be seen in these figures, RSD_R results are similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and often smaller than the DTM of 0.5. On the other hand, RSD_F results are also similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and often close to the DTM of 0.5. For DIF in new form only condition, there is an increase in RSD values for all methods. The RSD_R results of all methods except the CE equating method are close to the DTM of 0.5. The results of the CE equating method are smaller for a specific score range. However, this method gives quite high RSD_R values in the score ranges with low frequencies. The RSD_F results of all methods except the CE equating method are close to the DTM of 1. The results of CE equating method are smaller for a specific score range. However, this method gives quite high RSD_F values in the score ranges with low frequencies. To sum up, according to the results shown in Figures 25, and 26, the method that behaves most similarly to the criterion equating relationship in terms of both distributions and values is the SMO.

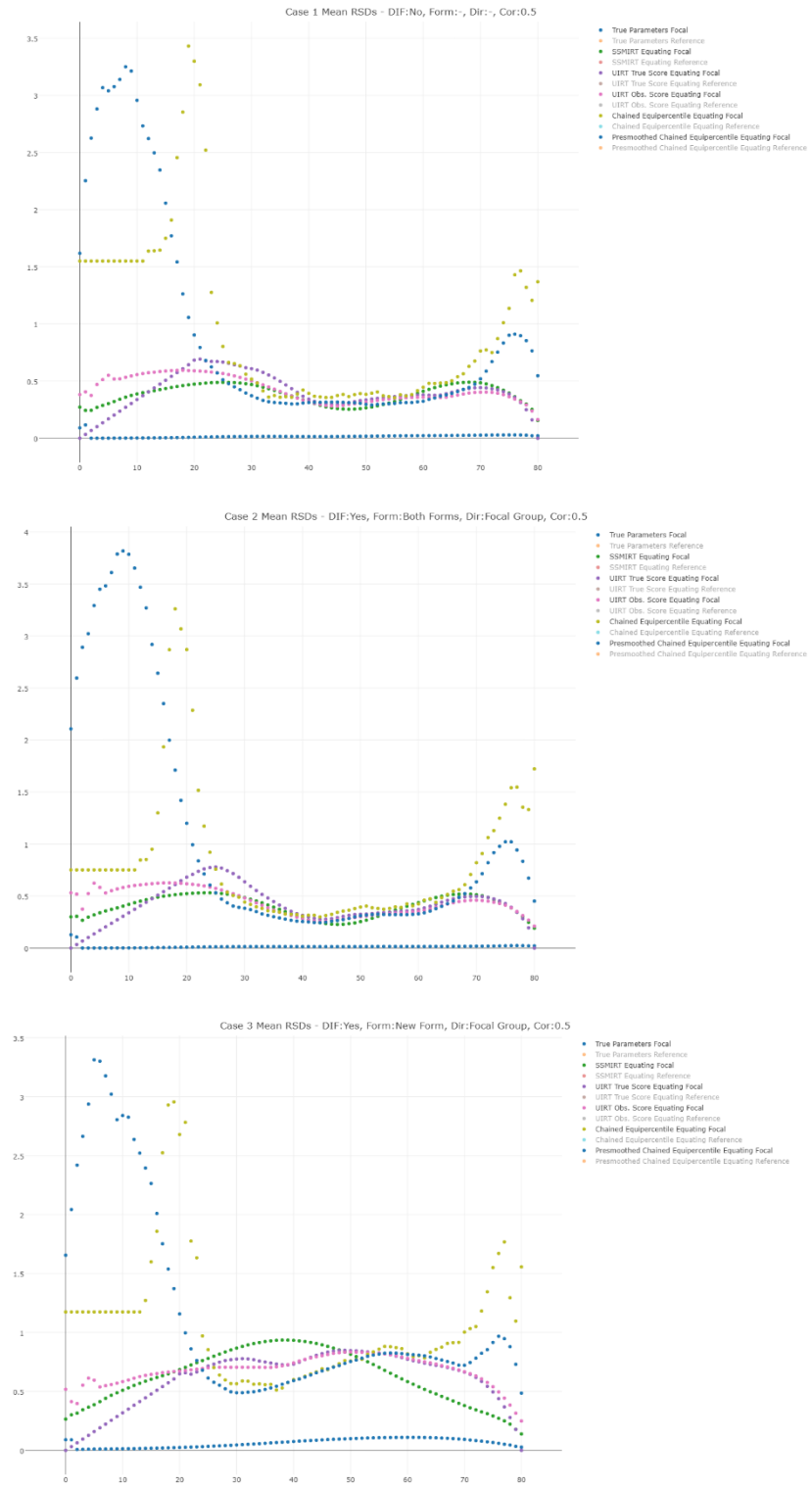

Figure 25. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

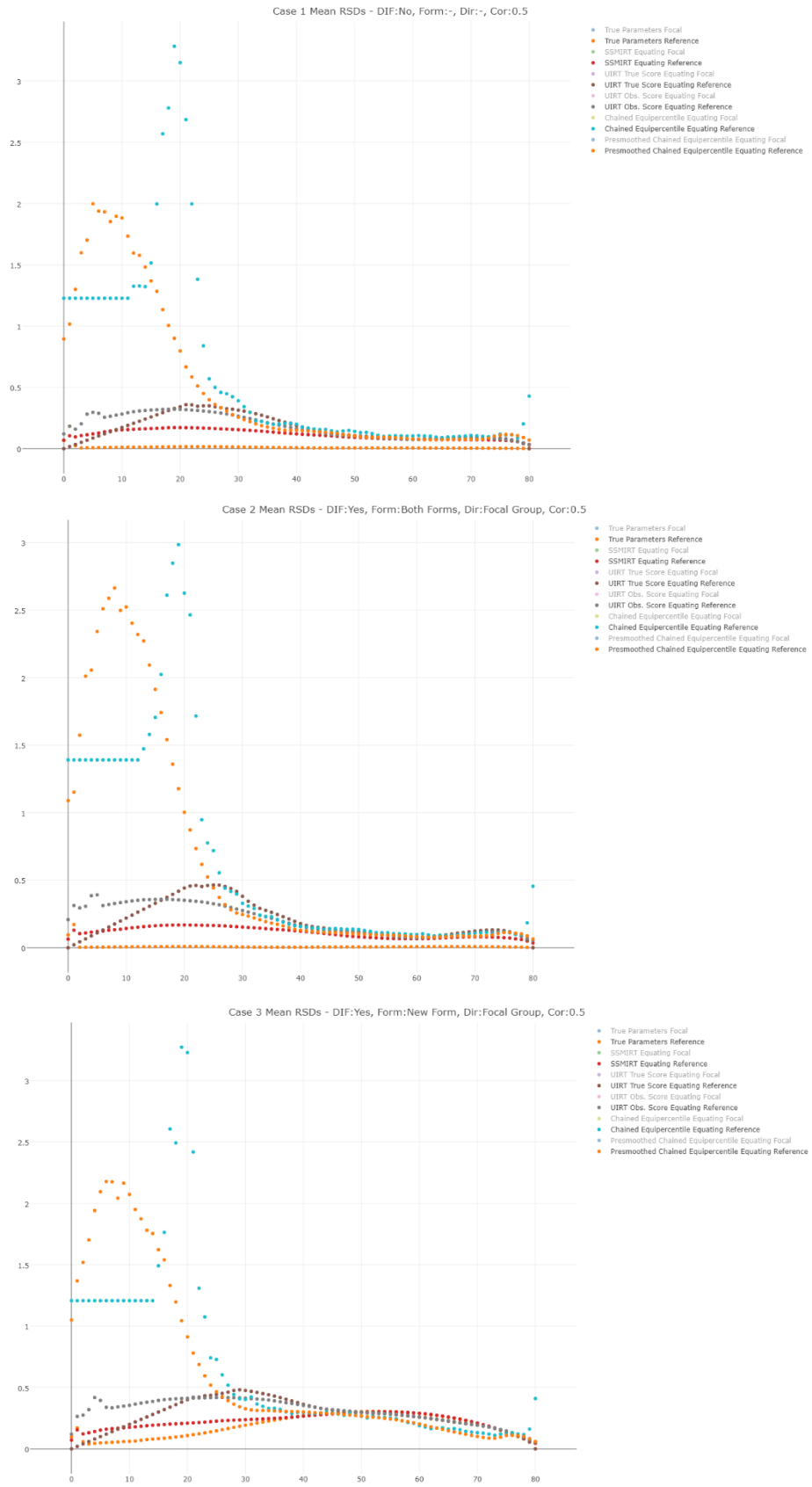

Figure 26. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.1.

Figures 27, and 28 represent the RSD_F and RSD_R means, respectively. These figures include no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation between dimensions and ES=0.3.

Figure 27. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.3.

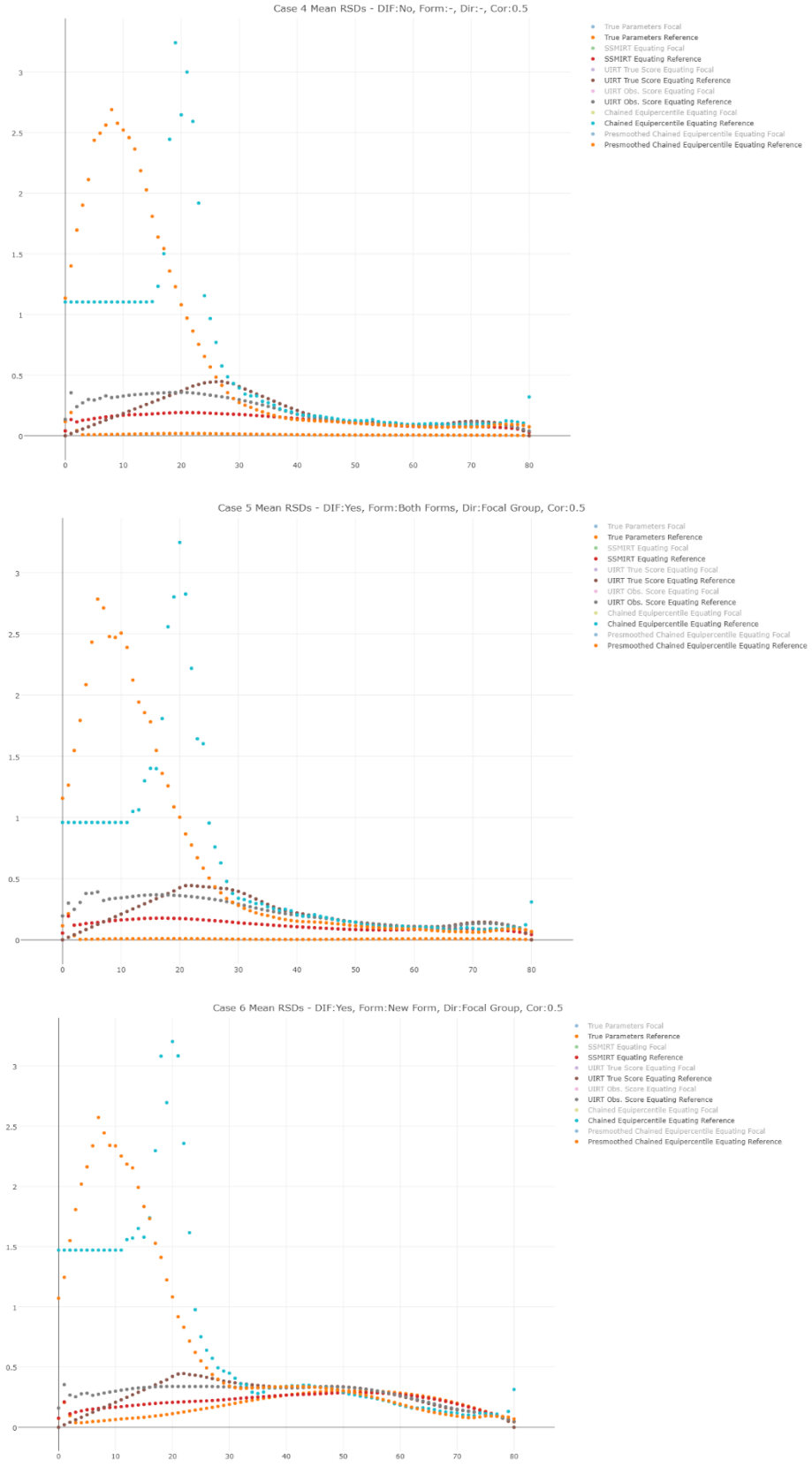

Figure 28. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.5 correlation and ES=0.3.

According to Figures 27, and 28, the results are quite similar to those for ES = 0.1. Specifically, RSD_R results are similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and often smaller than the DTM of 0.5. On the other hand, RSD_F results are also similar across no-DIF and DIF in both forms conditions, smaller than the DTM of 1.0, and often close to the DTM of 0.5. For DIF in new form only condition, there is an increase in RSD values for all methods. Specifically, the RSD_R results of all methods except the CE equating method are around the DTM of 0.5. The results of CE equating method are smaller for a specific score range however, this method gives quite high RSD_R values in the score ranges with low frequencies. Besides, the RSD_F results of all methods except the CE equating method are around the DTM of 1. The results of CE equating method are smaller for a specific score range however, this method gives also quite high RSD_F values in the score ranges with low frequencies. To sum up, according to the results shown in Figures 27, and 28, SMO equating method gives the closest results to the criterion equating relationship in terms of the distribution and the values. Based on these results mentioned for ES=0.3, it can be said that difference in ES has not an impact on the distribution and the values of RSD results. Consequently, in the conditions mentioned, the equating method that works closest to the criterion equating relationship in terms of the distribution and the values is the SMO.

RSD_F, and RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation, and ES=0.1 are represented in Figures 29 and 30 as below.

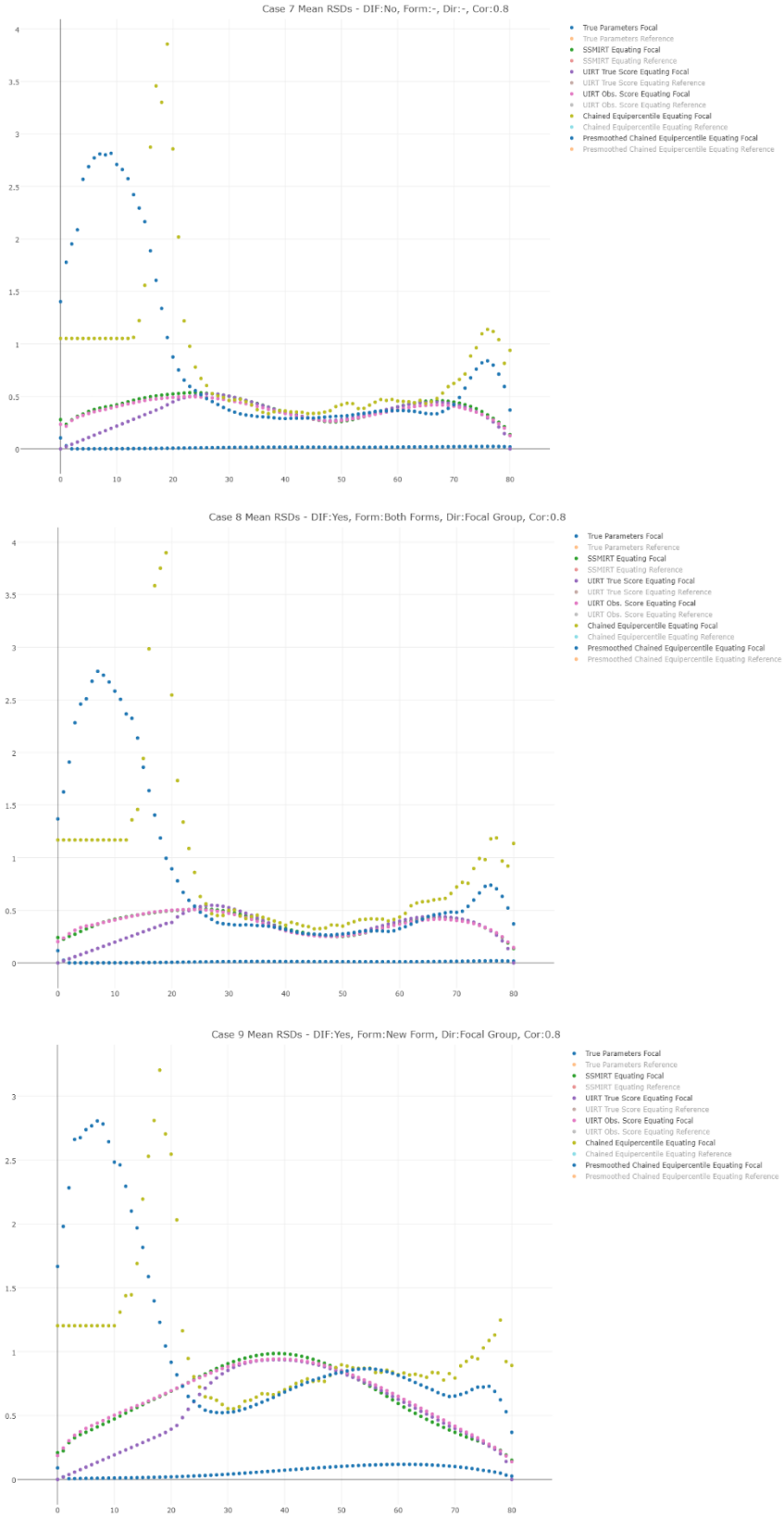

Figure 29. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.1.

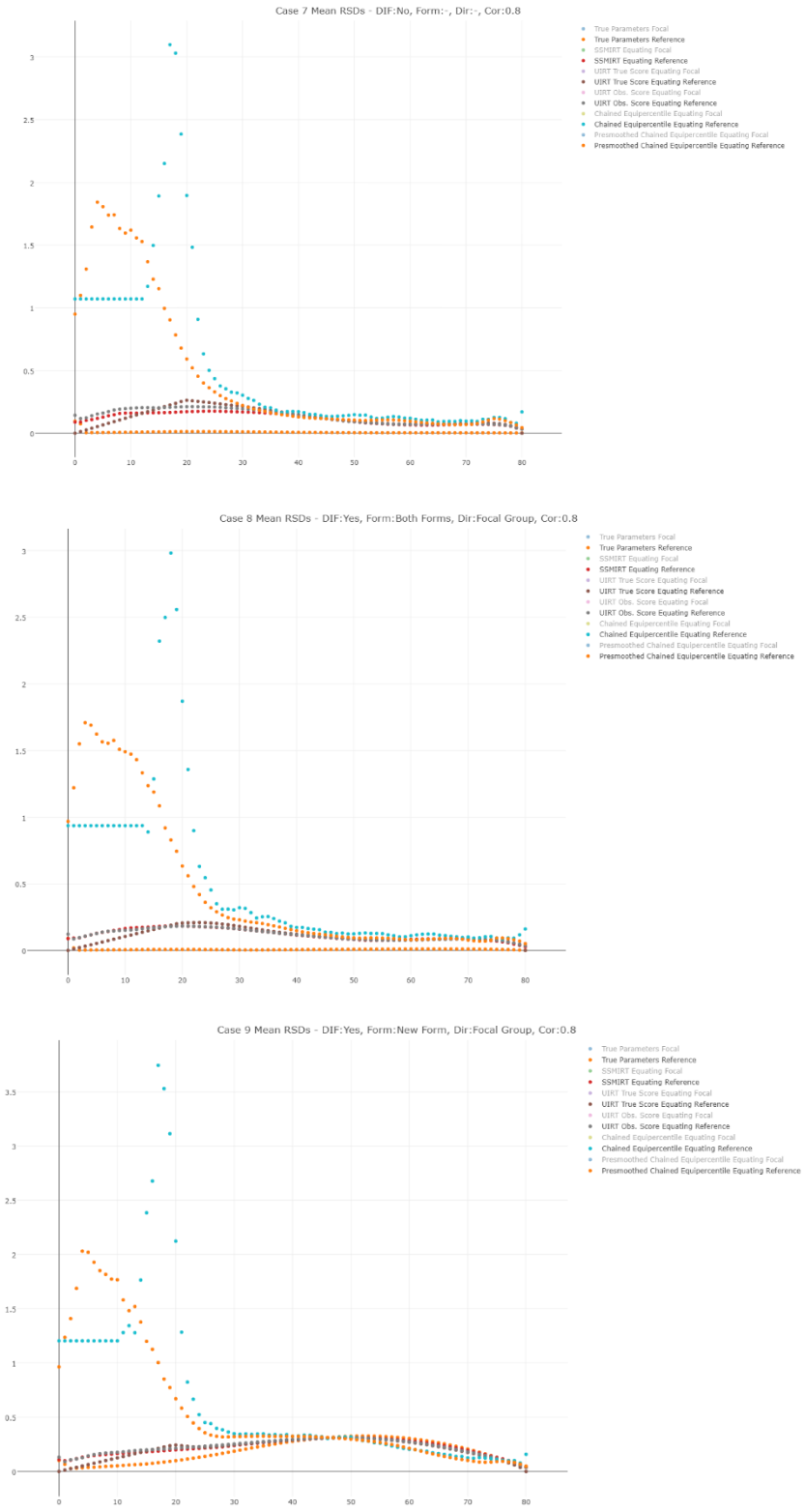

Figure 30. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.1.

According to Figures 29 and 30, with 0.8 correlation between dimensions and ES=0.1, the comparison of all equating methods according to no-DIF, DIF in both forms and DIF in new form only conditions is as follows. For no-DIF and DIF in both forms conditions, the RSD_R results of MIRT and UIRT methods, and the RSD_R results of the EQ method for scores with high frequencies are below the DTM of 0.5. On the other hand, for no-DIF and DIF in both forms conditions, the RSD_F results of MIRT and UIRT methods, and the RSD_F results of the EQ method for scores with high frequencies are below the DTM of 1, and often close to the DTM of 0.5. When DIF is added to new form only, the RSD results increase, and RSD_F results reach around the DTM of 1. That is, DIF in one form only causes an increase in the RSD values of all methods. The results obtained from the conditions mentioned are quite similar in terms of MIRT and UIRT equating methods. In some score ranges SMO, in some others UT, and in some others UO method gives results closer to the criteria. Consequently, in general it seems difficult to distinguish all methods in terms of the distributions and values of the RSD results.

RSD_F, and RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation, and ES=0.3 are represented in Figures 31 and 32, respectively. As can be seen in these figures, the RSD distributions of the methods are similar to those for conditions with ES=0.1. Specifically, for no-DIF and DIF in both forms conditions, the RSD_R results of MIRT and UIRT methods, and the RSD_R results of the EQ method for scores with high frequencies are below the DTM of 0.5. On the other hand, for no-DIF and DIF in both forms conditions, the RSD_F results of MIRT and UIRT methods, and the RSD_F results of the EQ method for scores with high frequencies are below the DTM of 1, and often close to the DTM of 0.5. When DIF is added to new form only, the RSD results increase, and RSD_F results reach around the DTM of 1. Besides, MIRT and UIRT equating RSD results are quite similar in terms of the values and the distributions. In general, it can be emphasized that all methods behave similarly for 0.8 correlation with ES=0.3.

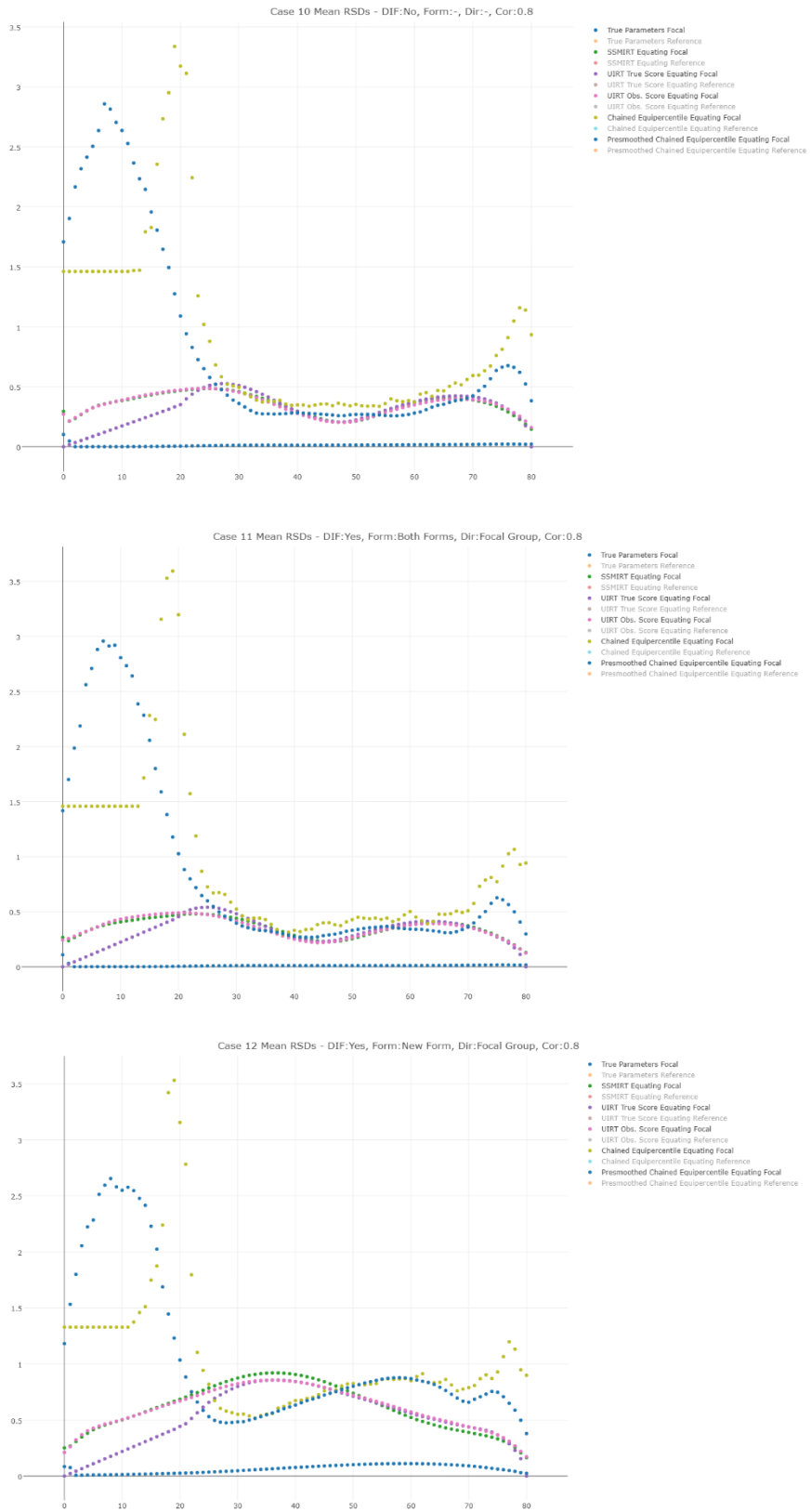

Figure 31. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3.

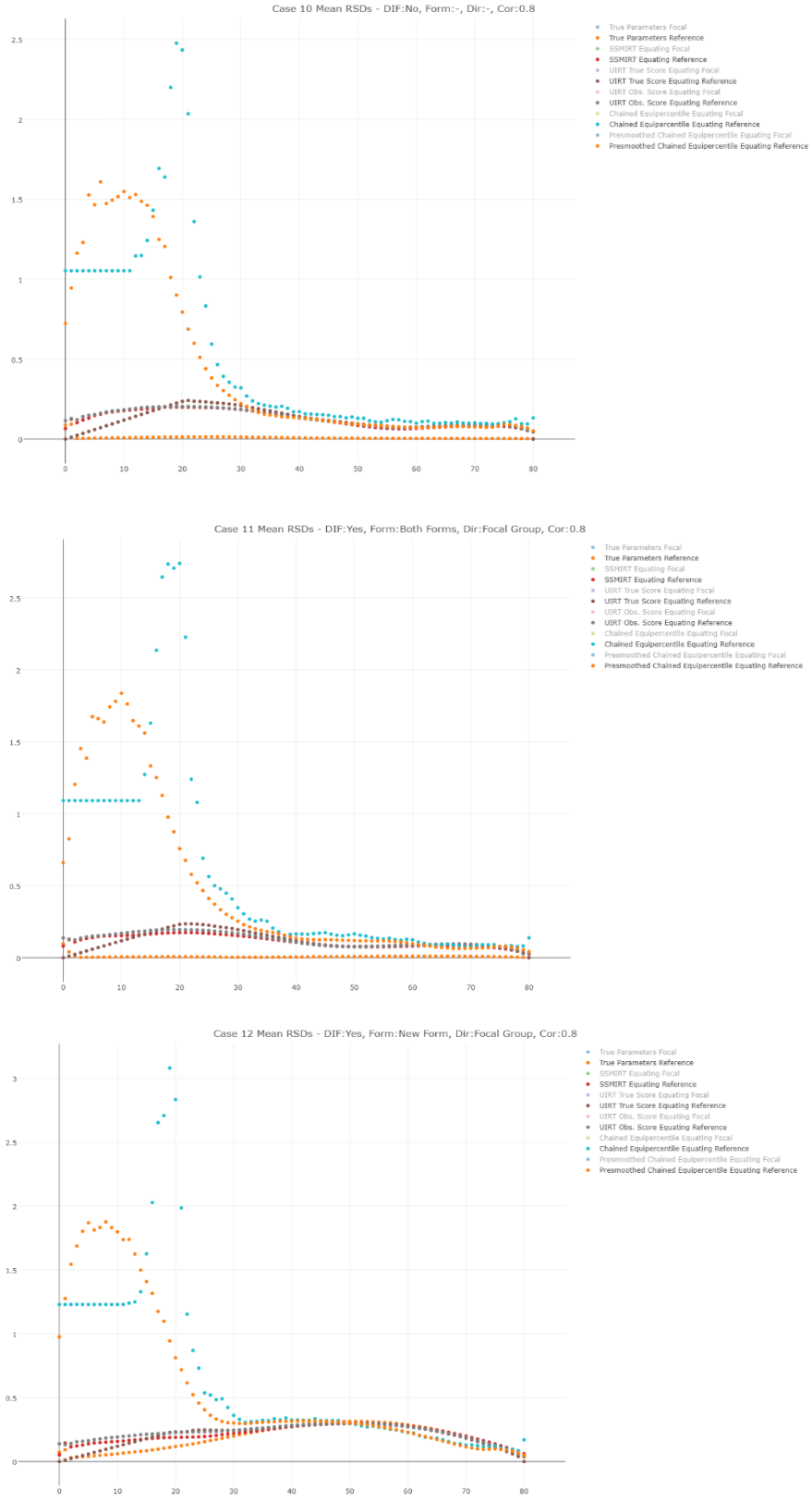

Figure 32. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.8 correlation and ES=0.3.

Figures 33 and 34 show the RSD_F and RSD_R means, respectively. These figures include no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation between dimensions and ES=0.1.

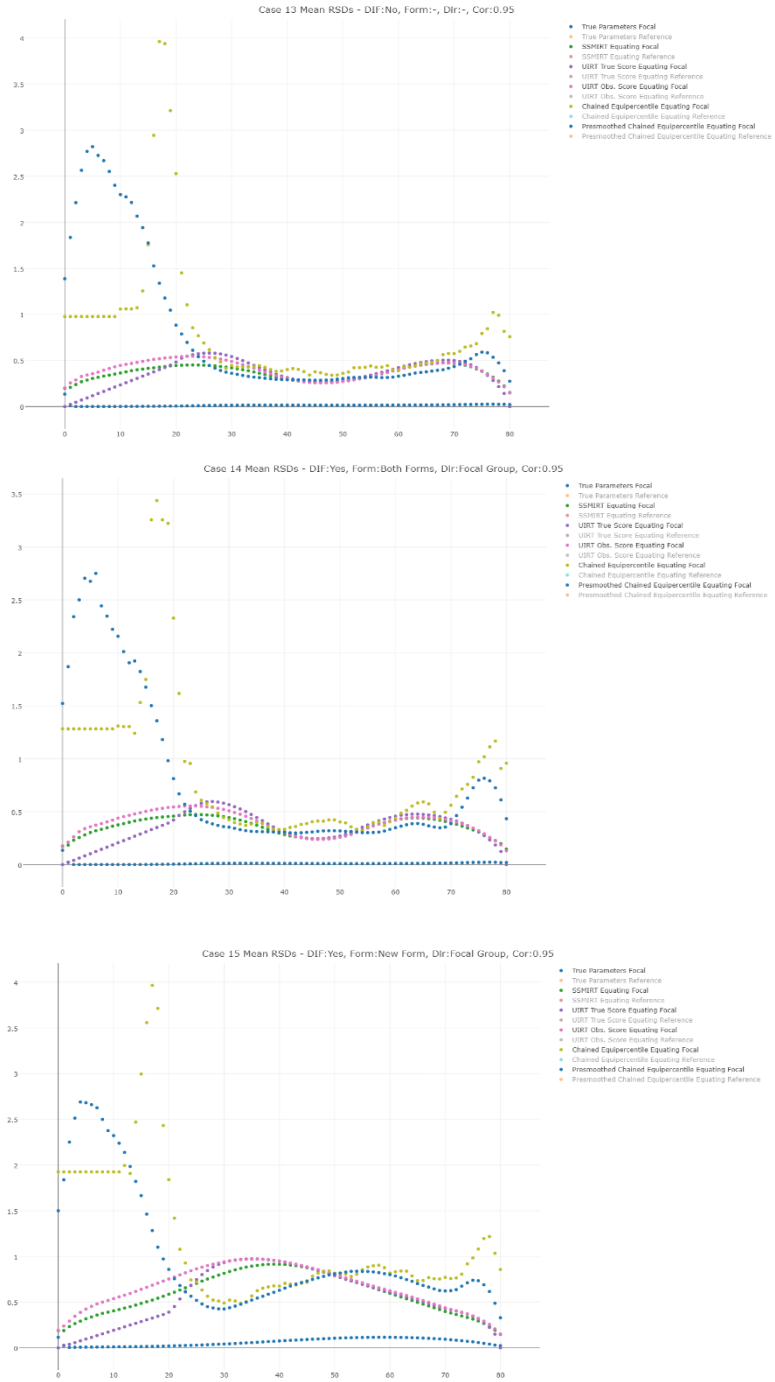

Figure 33. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.1.

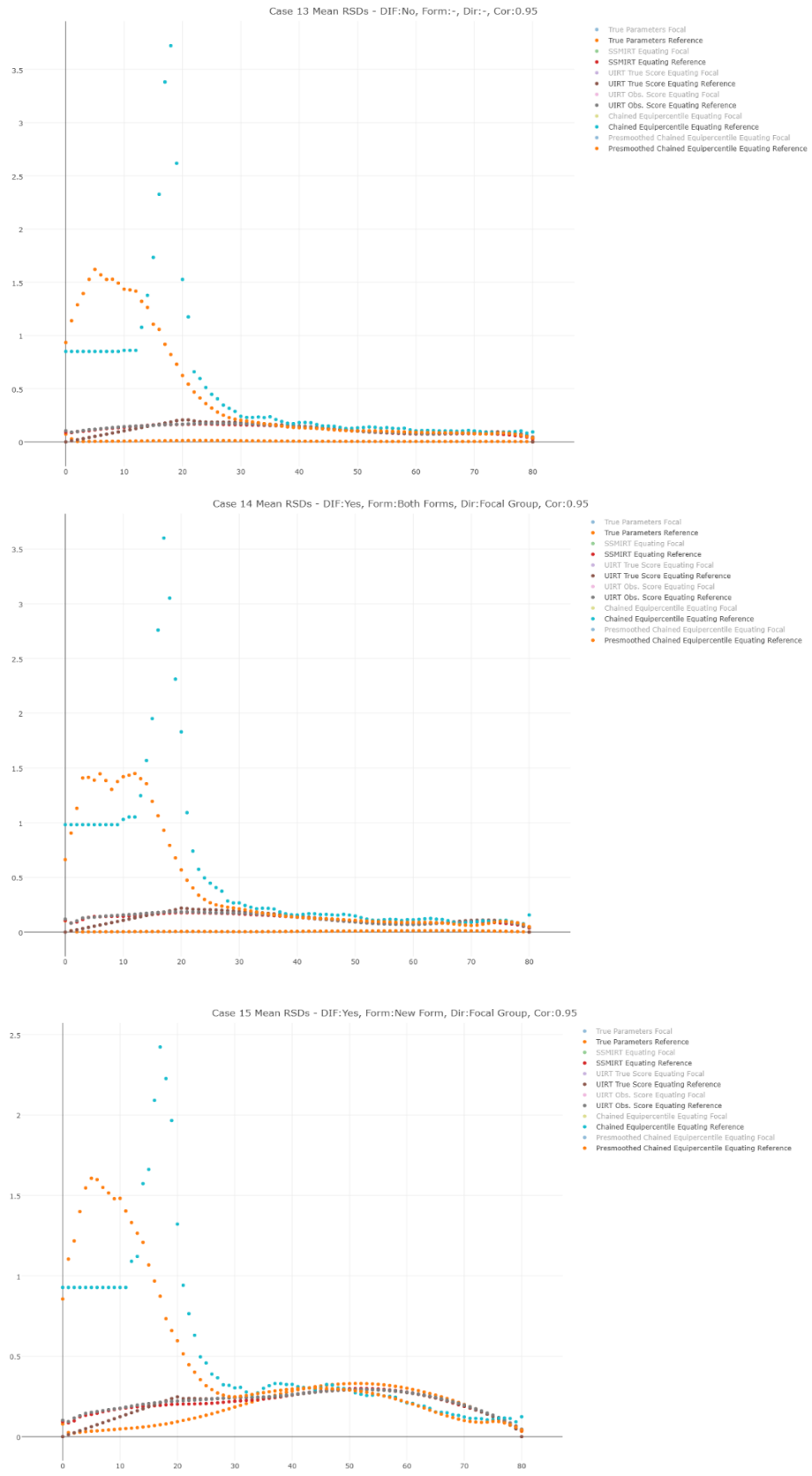

Figure 34. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.1.

Findings with respect to Figures 33 and 34 are as follows. With 0.95 correlation between dimensions and ES=0.1, for no-DIF and DIF in both forms conditions, the RSD_R results of MIRT and UIRT methods, and the RSD_R results of the EQ method for scores with high frequencies are smaller than the DTM of 0.5. Additionally, for no-DIF and DIF in both forms conditions, the RSD_F results of MIRT and UIRT methods, and the RSD_F results of the EQ method for scores with high frequencies are below the DTM of 1, and often close to the DTM of 0.5. When DIF is added to new form only, the RSD results increase for all methods. For the cases mentioned, the results are similar in terms of MIRT and UIRT methods and the EQ method for the scores with high frequencies.

When ES changes to 0.3, RSD distributions obtained are given in Figures 35 and 36 as below. According to findings shown in these figures, RSD results obtained are similar to those for conditions with ES=0.1. Specifically, with 0.95 correlation between dimensions and ES=0.3, for no-DIF and DIF in both forms conditions, the RSD_R results of MIRT and UIRT methods and results of the EQ method for scores with high frequencies are quite smaller than the DTM of 0.5. Additionally, for no-DIF and DIF in both forms conditions, the RSD_F results of MIRT and UIRT methods, and the RSD_F results of the EQ method for scores with high frequencies are below the DTM of 1, and often close to the DTM of 0.5. When DIF is added to new form only, all RSD results increase. Consequently, MIRT and UIRT results, and the EQ method results for scores with high frequencies are similar to each other as in the 0.8 correlation condition.

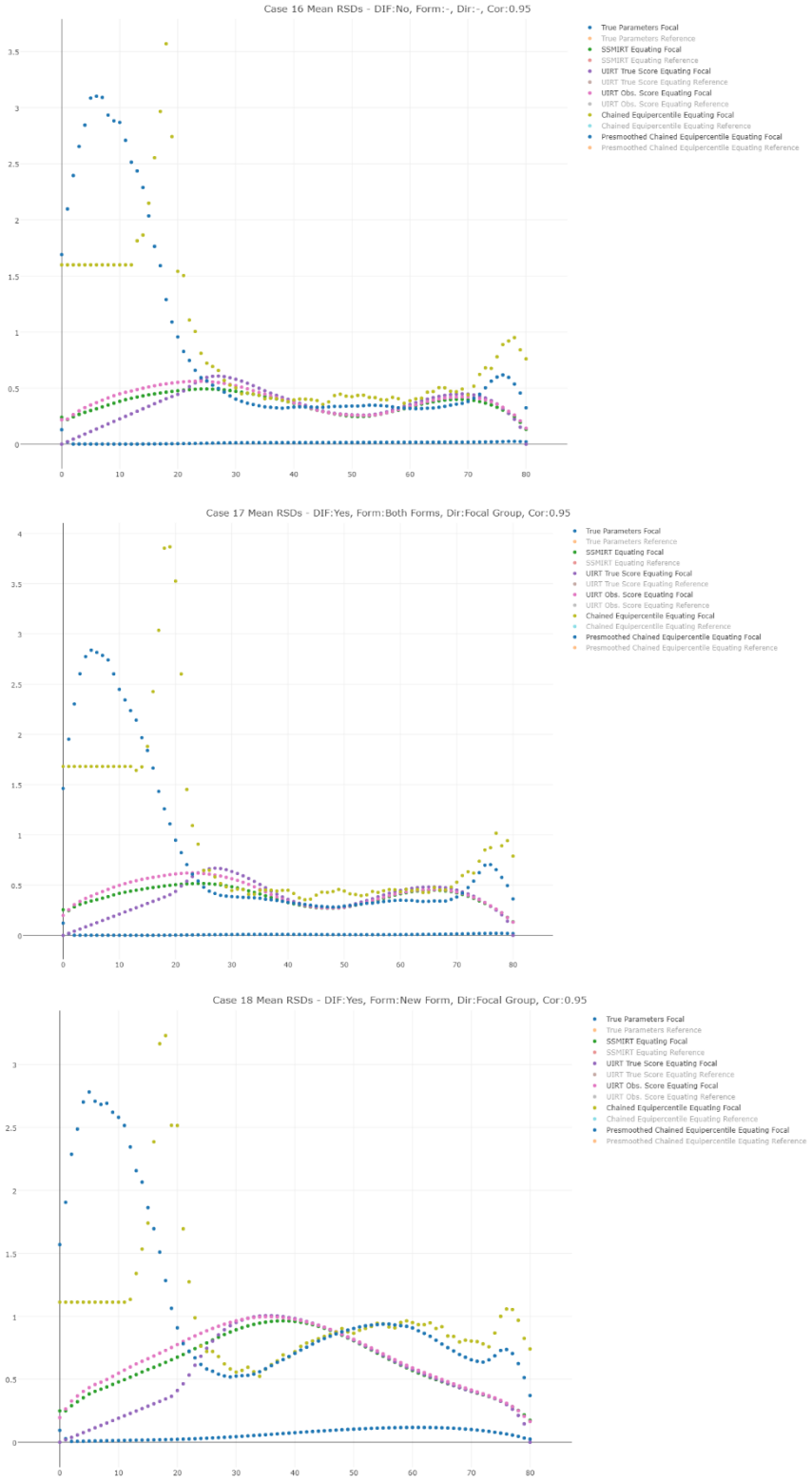

Figure 35. RSD_F means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.3.

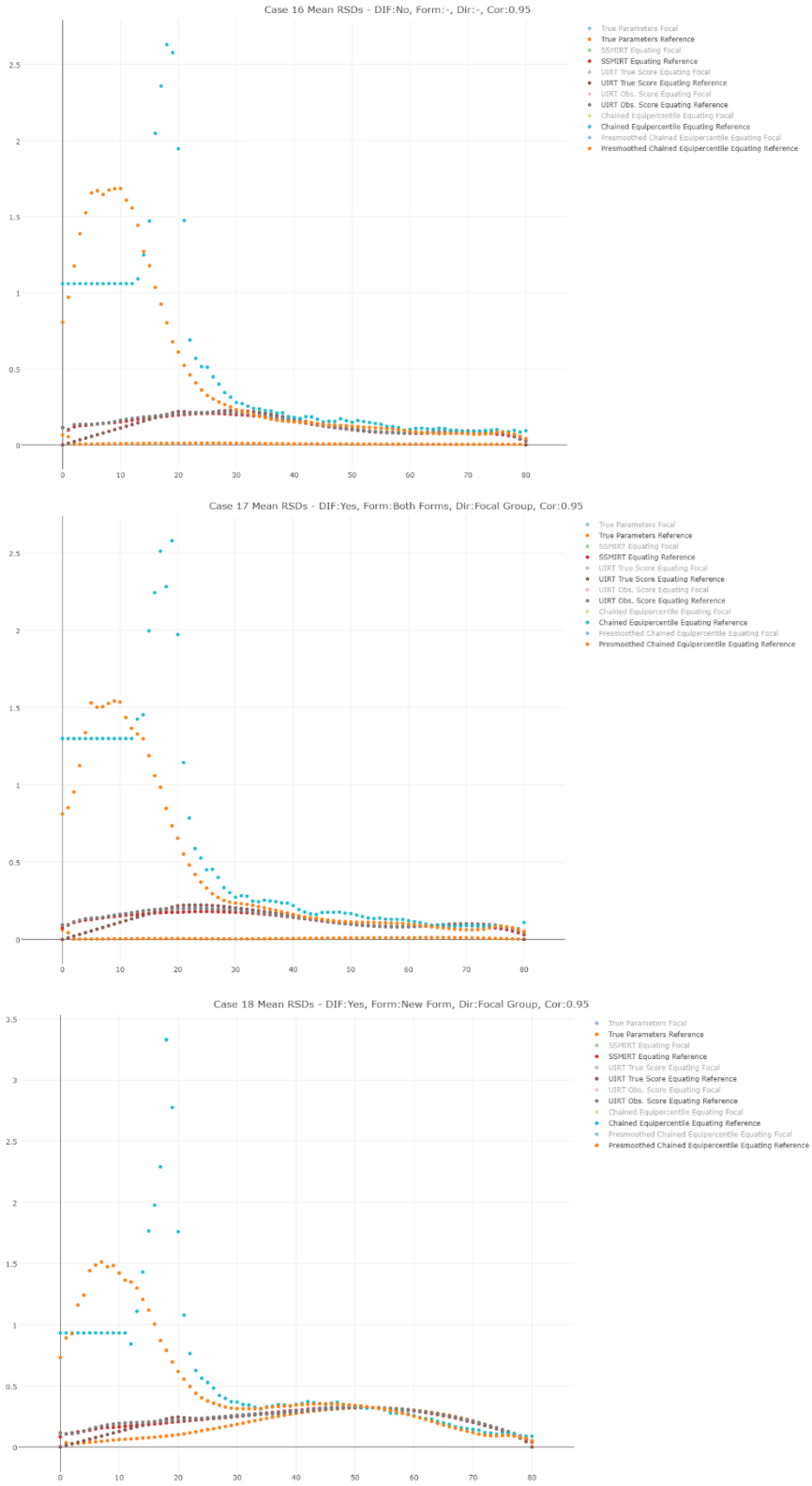

Figure 36. RSD_R means for no-DIF, DIF in both forms, and DIF in new form only conditions with 0.95 correlation and ES=0.3.

As can be seen in above figures, group-to-overall conditional equating invariance (the RSD_F, and the RSD_R) results for differential form DIF conditions show the same patterns as the RMSD results. However, for the RSD_F and the RSD_R results, it should be noted that the magnitudes of equating dependence are larger for the focal group as compared to the reference group, and also the fluctuations of equating dependence across the scores are more dramatic. In general, the RSD_R results of the methods are quite close to the criteria for reference group while the RSD_F results are well above the criteria for focal group.

Research Question 2

What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of correlation between dimensions on population invariance?

Correlation between dimensions was formed as 0.5 from condition 1 to condition 6, 0.8 from condition 7 to condition 12, and 0.95 from condition 13 to condition 18. According to the findings of 0.5 correlation condition with no-DIF and ES=0.1 (Case 1 – in Figures 7, 8 and 9), all methods except SMO give high RMSD results, close to the DTM of 0.5. Unlike the others, SMO gives the closest results to the criterion equating relationship in terms of the distribution and values. Besides, for the score rage with high frequencies CE equating method also gives close results to the criterion equating relationship, however for the score ranges with low frequencies the results of this method are quite high. For 0.8 correlation condition with no-DIF and ES=0.1 (Case 7 – in Figures 13, 14 and 15), all methods give close RMSD results to each other. Also, the RMSD results of all methods are smaller than the DTM of 0.5, except the CE equating results for the score ranges with low frequencies. For 0.95 correlation condition with no-DIF and ES=0.1 (Case 13 – in Figures 19, 20 and 21), the RMSD results of all methods are smaller than the DTM of 0.5 and are quite similar, except the CE equating method's results with low frequencies.

For 0.5 correlation condition with DIF in both forms and ES=0.1 (Case 2 – in Figures 7, 8 and 9), UT gives the highest RMSD results compared to SMO and UO methods. The RMSD results of this method are slightly above the DTM of 0.5. On the other hand, UO results are around the DTM of 0.5. Unlike other methods, the

RMSD results of SMO are well below the DTM of 0.5 and hence considerably lower than the results of the other methods. Again, for the score ranges with high frequencies the CE equating method gives very close results to the SMO and the criterion equating relationship results. For 0.8 correlation condition with DIF in both forms and ES=0.1 (Case 8 – in Figures 13, 14 and 15), the RMSD results of SMO, UO, and UT methods are close to each other, and below the DTM of 0.5. For the CE equating method, the RMSD results are below the DTM of 0.5 at the score range with high frequencies, and are very close to the other methods' results at this score range. For 0.95 correlation condition with DIF in both forms and ES=0.1 (Case 14 – in Figures 19, 20 and 21), except for the results of the CE equating method in low frequency score ranges, all methods yield results close to each other and below the DTM of 0.5.

At the 0.5 correlation condition with DIF in new form only and ES=0.1 (Case 3 – in Figures 7, 8 and 9), IRT methods give RMSD results above the DTM of 0.5. The UT method has the highest RMSD results among the SMO, UO and UT methods. It is followed by the UO method. On the other hand, the SMO method gives the closest result to the criterion equating relationship in terms of both the distribution and values. At the score range with high frequencies, the CE equating method's results are also close to the criterion equating relationship results. However, this method's results are interpreted separately from other methods due to the high RMSD results at the score ranges with low frequencies. For 0.8 correlation condition with DIF in new form only and ES=0.1 (Case 9 – in Figures 13, 14 and 15), the results of SMO, UO, and UT methods are very close to each other, and around the DTM of 0.5. It should be noted that the CE equating method's results are very close to the criterion equating relationship results, however, just for the scores with high frequencies. For the 0.95 correlation condition (Case 15 – in Figures 19, 20 and 21), results look similar to those of 0.8. Specifically, SMO, UO, and UT give very close results to each other, around the DTM of 0.5. On the other hand, the CE equating method's results are very close to the criterion equating relationship results but just for the scores with high frequencies.

When ES = 0.3, the results appear to be similar to those in ES=0.1. Specifically, for 0.5 correlation condition with no-DIF and ES=0.3 (Case 4 – in Figures 10, 11 and 12), all methods except SMO give high RMSD results, close to

the DTM of 0.5. Unlike the others, SMO gives the closest results to the criterion equating relationship in terms of the distribution and values. Besides, for the score rage with high frequencies CE equating method also gives close results to the criterion equating relationship, however for the score ranges with low frequencies the results of this method are quite high. For 0.8 correlation condition with no-DIF and ES=0.3 (Case 10 – in Figures 16, 17 and 18), all methods give close RMSD results to each other. Also, the RMSD results of all methods are smaller than the DTM of 0.5, except the CE equating results for the score ranges with low frequencies. For 0.95 correlation condition with no-DIF and ES=0.3 (Case 16 – in Figures 22, 23 and 24), The RMSD results of all methods are smaller than the DTM of 0.5 and are similar, except the CE equating method's results with low frequencies. To sum up, for 0.5 correlation condition, the methods which give the closest results to the criterion equating relationship are the SMO method, and the CE equating method for high frequencies. For the correlations of 0.8 and 0.95, the results of all methods, except for the EQ equating results in the scores with low frequencies, are close to each other and to the criterion equating relationship.

For 0.5 correlation condition with DIF in both forms and ES=0.3 (Case 5 – in Figures 10, 11 and 12), UT gives the highest RMSD results compared to SMO and UO methods. The RMSD results of this method are slightly above the DTM of 0.5. On the other hand, UO results are around the DTM of 0.5. Unlike other methods, the RMSD results of SMO are well below the DTM of 0.5 and hence considerably lower than the results of the other methods. Again, for the score ranges with high frequencies the CE equating method gives very close results to the SMO and the criterion equating relationship results. On the other hand, for 0.8 correlation condition with DIF in both forms and ES=0.3 (Case 11 – in Figures 16, 17 and 18), the RMSD results of SMO, UO, and UT methods are close to each other, and below the DTM of 0.5. For the CE equating method, the RMSD results are below the DTM of 0.5 at the score range with high frequencies, and are very close to the other methods' results at this score range. For 0.95 correlation condition with DIF in both forms and ES=0.3 (Case 17 – in Figures 22, 23 and 24), except for the results of the CE equating method in low frequency score ranges, all methods yield results close to each other and below the DTM of 0.5.

At the 0.5 correlation condition with DIF in new form only and ES=0.3 (Case 6 – in Figures 10, 11 and 12), the UT method has the highest RMSD results among the SMO, UO and UT methods. On the other hand, the SMO method gives the closest result to the criterion equating relationship in terms of both the distribution and values. At the score range with high frequencies, the CE equating method's results are also close to the criterion equating relationship results. However, these method's results are quite high at the score ranges with low frequencies. For 0.8 correlation condition with DIF in new form only and ES=0.3 (Case 12 – in Figures 16, 17 and 18), the results of SMO, UO, and UT methods are very close to each other, and around the DTM of 0.5. It should be noted that the CE equating method's results are also close to the criterion equating relationship results, however, just for the scores with high frequencies. For the 0.95 correlation condition (Case 18 – in Figures 22, 23 and 24), results look similar to those of 0.8. Specifically, SMO, UO, and UT give very close results to each other, around the DTM of 0.5. On the other hand, the CE equating method's results are close to IRT methods results but just for the scores with high frequencies.

The results of the other evaluation index discussed in the study are as follows. According to the RSD_F distributions of 0.5 correlation condition with no-DIF and ES=0.1 (Case 1 – in Figure 25), among the IRT methods SMO gives the lowest results. And also, SMO gives the closest results to the criterion equating relationship in terms of the distribution and values. Besides, for the score rage with high frequencies CE equating method also gives close results to the criterion equating relationship, however for the score ranges with low frequencies the results of this method are quite high. On the other hand, according to the RSD_R distributions of 0.5 correlation condition with no-DIF and ES=0.1 (Case 1 – in Figure 26), among all methods SMO gives the lowest results. So, SMO method results are the closest to the criterion equating relationship results in terms of the distribution and values. For 0.8 correlation condition with no-DIF and ES=0.1 (Case 7 – in Figure 29), all methods give close RSD_F results to each other. Also, the RSD_F results of all methods are smaller than the DTM of 1, and close to the DTM of 0.5, except the CE equating results for the score ranges with low frequencies. Additionally, all methods give close RSD_R results to each other (Case 7 – in Figure 30). And, the RSD_R results of all methods are quite smaller than the DTM of 0.5, except the CE equating results

for the score range with low frequencies. For 0.95 correlation condition with no-DIF and ES=0.1 (Case 13 – in Figure 33), The RSD_F results of all methods are smaller than the DTM of 1, and are around the DTM of 0.5. Also, all methods' RSD_F results are quite similar, except the CE equating method's results with low frequencies. On the other hand, except the CE equating method's results for the scores with low frequencies, the RSD_R results of all methods are quite smaller than the DTM of 0.5, and are quite similar to each other (Case 13 – in Figure 34).

For 0.5 correlation condition with DIF in both forms and ES=0.1 (Case 2 – in Figure 25), UT gives the highest RSD_F results compared to SMO and UO methods. The RSD_F results of this method are above the DTM of 0.5. The RSD_F results of the SMO are lower than the results of the other methods. Also, for the score ranges with high frequencies the CE equating method gives very close results to the SMO and the criterion equating relationship results. On the other hand, for 0.5 correlation condition with DIF in both forms and ES=0.1 (Case 2 – in Figure 26), UT gives the highest RSD_R results compared to SMO and UO methods. The RSD_R results of this method are close to the DTM of 0.5. The RSD_R results of the SMO are lower than the results of the other methods. Among all methods, the SMO is the method which gives the closest results to the criterion equating relationship. For 0.8 correlation condition with DIF in both forms and ES=0.1 (Case 8 – in Figure 29), the RSD_F results of SMO, UO, and UT methods are close to each other, and are around the DTM of 0.5. For the CE equating method, the RSD_F results are also around the DTM of 0.5 at the score range with high frequencies, and are very close to the other methods' results at this score range. Additionally, for 0.8 correlation condition with DIF in both forms and ES=0.1 (Case 8 – in Figure 30), the RSD_R results of SMO, UO, and UT methods are close to each other, and below the DTM of 0.5. For the CE equating method, the RSD_R results are also around the DTM of 0.5 at the score range with high frequencies, and are very close to the other methods' results at this score range. For 0.95 correlation condition with DIF in both forms and ES=0.1 (Case 14 – in Figure 33), except for the results of the CE equating method at the score ranges with low frequencies, all methods yield RSD_F results close to each other and around the DTM of 0.5. On the other hand, the RSD_R results of all methods (Case 14 – in Figure 34), except the CE equating method results at the scores with low frequencies, are close to each other, and are below the DTM of 0.5.

At the 0.5 correlation condition with DIF in new form only and ES=0.1 (Case 3 – in Figure 25), all methods give RSD_F results above the DTM of 0.5. On the other hand, the SMO method behaves similarly to the criterion equating relationship in terms of both the distribution and values. At the score range with high frequencies, the CE equating method's results are also close to the criterion equating relationship results. However, this method's results are interpreted separately from other methods due to the high RSD_F results at the score ranges with low frequencies. Additionally, the RSD_R results of the SMO method are smaller than the UO and UT methods results (Case 3 – in Figure 26). And, among all methods the SMO method gives closest results to the criterion equating relationship results. For 0.8 correlation condition with DIF in new form only and ES=0.1 (Case 9 – in Figure 29), the RSD_F results of SMO, UO, and UT methods are very close to each other, and are around the DTM of 1. It should be noted that the CE equating method's results are close to the other methods results, however, just for the scores with high frequencies. The RSD_R results of all methods are also close to each other, and are below the DTM of 0.5 (Case 9 – in Figure 30). For the 0.95 correlation condition (Case 15 – in Figure 33), results look similar to those of 0.8. Specifically, all methods, except the CE equating method at the scores with low frequencies, give very close RSD_F results to each other, and these values are around the DTM of 1. Also, the RSD_R results of all methods, except the CE equating method results at the scores with low frequencies, are close to each other, and are below the DTM of 0.5 (Case 15 – in Figure 34).

When ES = 0.3, the RSD results appear to be similar to those in ES=0.1. Specifically, for 0.5 correlation condition with no-DIF and ES=0.3 (Case 4 – in Figure 27), among IRT methods, SMO gives the closest RSD_F results to the criterion equating relationship in terms of both the distribution and values. Besides, for the score rage with high frequencies CE equating method also gives close results to the criterion equating relationship, however for the score ranges with low frequencies the results of this method are quite high. On the other hand, according to the RSD_R results (Case 4 – in Figure 28), the method which gives the closest results to the criterion equating relationship is the SMO. For 0.8 correlation condition with no-DIF and ES=0.3 (Case 10 – in Figure 31), all methods give close RSD_F results to each other. Also, the RSD_F results of all methods are smaller than the DTM of 1, and are

around the DTM of 0.5, except the CE equating results for the score ranges with low frequencies. The RSD_R results of all methods are well below the DTM of 0.5, and are close to each other, except the CE equating results for the score ranges with low frequencies (Case 10 – in Figure 32). For 0.95 correlation condition with no-DIF and ES=0.3 (Case 16 – in Figure 35), except the CE equating method's results for low frequencies, the RSD_F results of all methods are around the DTM of 0.5 and are similar to each other. As in the 0.8 correlation condition, the RSD_R results of all methods are well below the DTM of 0.5, and are close to each other, except the CE equating results for the score ranges with low frequencies (Case 16 – in Figure 36). To sum up, for 0.5 correlation condition, the methods which give the closest results to the criterion equating relationship are the SMO method, and the CE equating method for high frequencies. For the correlations of 0.8 and 0.95, the results of all methods, except for the EQ method results at the scores with low frequencies, are close to each other.

For 0.5 correlation condition with DIF in both forms and ES=0.3 (Case 5 – in Figure 27), UT gives the highest RSD_F results compared to SMO and UO methods. Also, the SMO method gives the closest results to the criterion equating relationship results. Again, for the score ranges with high frequencies the CE equating method also gives very close results to the criterion equating relationship. For the conditions mentioned, UT gives the highest RSD_R results compared to SMO and UO methods (Case 5 – in Figure 28). And, the SMO method gives the closest RSD_R results to the criterion equating relationship results. On the other hand, for 0.8 correlation condition with DIF in both forms and ES=0.3 (Case 11 – in Figure 31), the RSD_F results of all methods, except the CE equating results at the score ranges with low frequencies, are very close to each other, and are around the DTM of 0.5. Additionally, the RSD_R results of all methods (Case 11 – in Figure 32), again except the CE equating results at the score ranges with low frequencies, are very close to each other, and are well below the DTM of 0.5. For 0.95 correlation condition with DIF in both forms and ES=0.3 (Case 17 – in Figure 35), except for the results of the CE equating method at the score ranges with low frequencies, all methods yield RSD_F results close to each other and around the DTM of 0.5. On the other hand, the RSD_R results of all methods (Case 17 – in Figure 36), again except the CE

equating results at the score ranges with low frequencies, are very close to each other, and are well below the DTM of 0.5.

At the 0.5 correlation condition with DIF in new form only and ES=0.3 (Case 6 – in Figure 27), based on the RSD_F results, the SMO method behaves most similarly to the criterion equating relationship in terms of both the distribution and values. At the score range with high frequencies, the CE equating method's results are also close to the criterion equating relationship results. However, these method's results are quite high at the score ranges with low frequencies. On the other hand, for RSD_R results the method which gives the closest results to the criterion equating relationship is again the SMO (Case 6 – in Figure 28). For 0.8 correlation condition with DIF in new form only and ES=0.3 (Case 12 – in Figure 31), the RSD_F results of SMO, UO, and UT methods are very close to each other, and around the DTM of 1. It should be noted that the CE equating method's results are also close to the criterion equating relationship results, however, just for the scores with high frequencies. Also, except the CE equating method results at the scores with low frequencies, the RSD_R results of all methods are very close to each other, and are well below the DTM of 0.5 (Case 12 – in Figure 32). For the 0.95 correlation condition with DIF in new form only and ES=0.3 (Case 18 – in Figure 35), RSD results look similar to those for 0.8. Specifically, SMO, UO, and UT give very close RSD_F results to each other, around the DTM of 1. On the other hand, the CE equating method's results at the scores with high frequencies are similar to IRT methods results. Again, except the CE equating method results at the scores with low frequencies, the RSD_R results of all methods are very close to each other, and are well below the DTM of 0.5 (Case 18 – in Figure 36).

Finally, it should be emphasized that the RSD_F, and the RSD_R results for all conditions of correlation between dimensions show the same patterns as the RMSD results. However, for the RSD_F, and the RSD_R results, the magnitudes of equating dependence are larger for the focal group as compared to the reference group, and also the fluctuations of equating dependence across the scores are more dramatic.

Research Question 3

What is the performance of MIRT equating method compared to UIRT and EQ equating methods with respect to the effect of group mean ability differences between two forms on the relationship between DIF and population invariance?

In this section, equating invariance results are interpreted with respect to group mean ability differences between the old and new forms. To form group ability difference between the forms, group ability means of both dimensions were increased as 0.1 and 0.3 unit in the new form. The same effect was created also in the focal and reference groups as in the total group. For example, while the focal group ability means for the old form were -0.3 for both dimensions, in the ES = 0.1 condition these means increased to -0.2 in the new form.

First, for ES=0.1 condition with no-DIF and 0.5 correlation (Case 1 – in Figures 7, 8 and 9), the RMSD results of the SMO, and the CE equating method at the score range with high frequencies are close to the criterion equating relationship results. UT method's results are higher than all methods, except the results of the CE equating method at the score ranges with low frequencies. Additionally, while the SMO method results, and the CE equating method results for the scores with high frequencies are well below the DTM of 0.5, UO and UT methods results reach around the DTM of 0.5. When ES changes to 0.3 (Case 4 – in Figures 10, 11 and 12), the results almost remain the same. That is, the SMO results, and the CE equating method results at the scores with high frequencies are close to the criterion relationship. Besides, UT method results are higher than the results of all methods, except the results of CE equating method at the score ranges with low frequencies. Also, while the SMO method results, and the CE equating method results for the scores with high frequencies are well below the DTM of 0.5, UO and UT methods results reach around the DTM of 0.5 and even exceed this value. For ES=0.1 with no-DIF and 0.8 correlation (Case 7 – in Figures 13, 14 and 15), MIRT and UIRT methods results are close to each other and well below the DTM of 0.5. On the other hand, the EQ method results are also below the DTM of 0.5 for the scores with high frequencies. When ES changes to 0.3 (Case 10 – in Figures 16, 17 and 18), the results again remain almost the same. Specifically, IRT methods results are close to each other, and are below the DTM of 0.5. Besides, the EQ method results are

also below the DTM of 0.5 for the scores with high frequency. And, in general all methods, except the CE equating method for the scores with low frequencies, behave similar. For ES=0.1 condition with no-DIF and 0.95 correlation (Case 13 – in Figures 19, 20 and 21), IRT methods' RMSD results are close to each other, and well below the DTM of 0.5. On the other hand, at the scores with high frequencies the EQ method's results are also well below the DTM of 0.5. It should be emphasized that all methods, except the CE equating method at the scores with low frequencies, behave similar, and give close results to the criterion equating relationship. For ES=0.3 (Case 16 – in Figures 22, 23 and 24), the results look similar to those for ES=0.1. That is, the RMSD means of IRT methods are close to each other, and below the DTM of 0.5. And, the EQ method results are also below the DTM of 0.5 at the scores with high frequencies. Consequently, for the conditions mentioned IRT methods, and the CE equating method (just for the scores with high frequencies) behave similarly to each other and to the criterion equating relationship.

Above, the effect of group mean ability differences on equating invariance is examined for no-DIF and various correlations. In this paragraph, the effect of group mean ability differences on equating invariance is examined for DIF in both forms and various correlations. For ES=0.1 condition with DIF in both forms and 0.5 correlation (Case 2 – in Figures 7, 8 and 9), UT method gives the highest RMSD results among all methods (except the CE equating method at the scores with low frequencies). Also, the methods which give the closest results to the criterion equating relationship are the SMO method, and the CE equating method for high frequencies. For ES=0.3 condition with DIF in both forms and 0.5 correlation (Case 5 – in Figures 10, 11 and 12), the RMSD results are similar to those for ES=0.1. That is, UT method gives the highest RMSD results, and the SMO method, and the CE equating method for high frequencies give the closest results to the criterion equating relationship. For ES=0.1 condition with DIF in both forms and 0.8 correlation (Case 8 – in Figures 13, 14 and 15), the results of all methods, except the CE equating method results for the scores with low frequencies, are very close to each other, and are below the DTM of 0.5. On the other hand, For ES=0.3 condition with DIF in both forms and 0.8 correlation (Case 11 – in Figures 16, 17 and 18), again results are similar to those for ES=0.1. Specifically, the results of all

methods, except the CE equating method results for the scores with low frequencies, are close to each other, and are below the DTM of 0.5. For ES=0.1 condition with DIF in both forms and 0.95 correlation (Case 14 – in Figures 19, 20 and 21), the RMSD means of all methods, again except the results of the CE equating method for the score ranges with low frequencies, are close to each other, and are below the DTM of 0.5. When ES changes to 0.3 (Case 17 – in Figures 22, 23 and 24), the RMSD results are almost the same. That is, the results of all methods, except the CE equating method results for the scores with low frequencies, are close to each other, and are below the DTM of 0.5. To sum up, the RMSD results of the methods are quite similar for 0.1 and 0.3 effect size conditions with DIF in both forms and various correlations.

For DIF in new form only, and various correlation conditions, the effects of group mean ability differences on equating invariance are as follows. For ES=0.1 condition with DIF in new form only and 0.5 correlation (Case 3 – in Figures 7, 8 and 9), the method which gives the closest results to the criterion equating relationship is the SMO. Also, the CE equating method results are also very close to the criterion equating relationship, however at the scores with low frequencies this method results are very high. On the other hand, the UT method gives the highest RMSD results among IRT methods. When ES changes to 0.3 (Case 6 – in Figures 10, 11 and 12), the results almost remain the same. That is, the SMO method, and the CE equating method for the scores with high frequencies behave similarly to the criteria. Additionally, unlike the other IRT methods results, the UT method results are above the DTM of 0.5. For ES=0.1 condition with DIF in new form only and 0.8 correlation (Case 9 – in Figures 13, 14 and 15), IRT methods results are very close to each other, and are slightly above the DTM of 0.5. The CE equating method results are also around the DTM of 0.5, but this is only valid for the scores with high frequencies. When ES changes to 0.3 (Case 12 – in Figures 16, 17 and 18), the results again remain the same. To mention in detail, all methods behave similarly, and the results of all methods, except again the CE equating method results for the scores with low frequencies, are around the DTM of 0.5. For ES=0.1 condition with DIF in new form only and 0.95 correlation (Case 15 – in Figures 19, 20 and 21), IRT methods results are very close to each other, and are slightly above the DTM of 0.5. Additionally, the CE equating method results, for the scores with high frequencies, are also close to

the DTM of 0.5. In other words, all methods, except the CE equating method for the scores with low frequencies, behave similarly for this case. When ES changes to 0.3 (Case 18 – in Figures 22, 23 and 24), the results remain the same. Specifically, all methods behave similarly, and the results of all methods, except again the CE equating method results for the scores with low frequencies, are slightly above the DTM of 0.5.

The results of the other evaluation index (RSD) discussed in the study are as follows. According to the findings, RSD results for all conditions with group mean ability difference between two forms show the same patterns as the RMSD results. First, for ES=0.1 condition with no-DIF and 0.5 correlation (Case 1 – in Figure 25), the RSDF results of the SMO, and the CE equating method at the score range with high frequencies are close to the criterion equating relationship results. UT method's results are higher than all methods, except the results of the CE equating method at the score ranges with low frequencies. On the other hand, based on RSDR results (Case 1 – in Figure 26), the method that gives the closest results to the criterion equating relationship is the SMO. And, the SMO method results are well below the DTM of 0.5. Also, all other methods' RSDR results, except the CE equating method results for the scores with low frequencies, are below the DTM of 0.5. For ES=0.3 condition with no-DIF and 0.5 correlation (Case 4 – in Figure 27), the RSD results almost remain the same. That is, the RSDF results of the SMO method, and the RSDF results of the CE equating method at the scores with high frequencies are closer to the criterion equating relationship than the results of other methods. Besides, UT method results are higher than the results of all methods, except the results of CE equating method at the score ranges with low frequencies. Based on RSDR results (Case 4 – in Figure 28), the method that gives the closest results to the criterion equating relationship is the SMO. And, the SMO method results are well below the DTM of 0.5 while the RSDR results of other methods, except the CE equating method results for the scores with low frequencies, are around the DTM of 0.5. For ES=0.1 condition with no-DIF and 0.8 correlation (Case 7 – in Figure 29), the RSDF results of all methods, except the CE equating method results for the scores with low frequencies, are close to each other, and are around the DTM of 0.5. On the other hand, the RSDR results of all methods, again except the CE equating method results for the scores with low frequencies, are close to each other,

and are well below the DTM of 0.5 (Case 7 – in Figure 30). When ES changes to 0.3 (Case 10 – in Figure 31), the RSD results again remain almost the same. To mention in detail, the RSDF results of all methods, except the CE equating method results for the score ranges with low frequencies, are close to each other and around the DTM of 0.5. Also, the RSDR results of all methods (Case 10 – in Figure 32), again except the CE equating method results at the scores with low frequencies, are very close to each other and well below the DTM of 0.5. For ES=0.1 condition with no-DIF and 0.95 correlation (Case 13 – in Figure 33), the RSDF results of all methods, except the CE equating method results at the score ranges with low frequencies, are close to each other, and are around the DTM of 0.5. According to the RSDR results (Case 13 – in Figure 34), again except the CE equating method at the scores with low frequencies, all methods behave similar, and give close results to the criterion equating relationship. When ES changes to 0.3 (Case 16 – in Figure 35), the results remain the same. That is, the RSDF means of all methods, except the CE equating method at the score ranges with low frequencies, are close to each other, and are around the DTM of 0.5. On the other hand, the RSDR results of all methods (Case 16 – in Figure 36), again except the CE equating method results at the scores with low frequencies, are very close to each other, and are well below the DTM of 0.5. Consequently, for the condition mentioned IRT methods, and the CE equating method (just for the scores with high frequencies) behave similarly to each other and to the criterion equating relationship.

Above, the effect of group mean ability differences on equating invariance is examined for no-DIF and various correlations. In this paragraph, the effect of group mean ability differences on equating invariance is examined for DIF in both forms and various correlations. For ES=0.1 condition with DIF in both forms and 0.5 correlation (Case 2 – in Figure 25), UT method gives the highest RSD_F results among all methods (except the CE equating method at the scores with low frequencies). The methods which give the closest results to the criterion equating relationship are the SMO method, and the CE equating method for high frequencies. Also, for the RSD_R results (Case 2 – in Figure 26), the SMO method gives the closest results to the criterion equating relationship. For ES=0.3 condition with DIF in both forms and 0.5 correlation (Case 5 – in Figure 27), the RSD_F results are similar to those for ES=0.1. That is, UT method gives the highest RSD_F results. And the SMO

method, and the CE equating method for high frequencies behave similarly to the criterion equating relationship. Additionally, also the RSD_R results are similar to those for ES=0.1 (Case 5 – in Figure 28). Specifically, the SMO method gives the closest results to the criterion equating relationship. For ES=0.1 condition with DIF in both forms and 0.8 correlation (Case 8 – in Figure 29), the RSD_F results of all methods, except the CE equating method results for the scores with low frequencies, are very close to each other, and are around the DTM of 0.5. And, the RSD_R results of all methods (Case 8 – in Figure 30), again except the CE equating method results for the scores with low frequencies, are very close to each other, and are well below the DTM of 0.5. On the other hand, For ES=0.3 condition with DIF in both forms and 0.8 correlation (Case 11 – in Figure 31), the RSD_F results are similar to those for ES=0.1. Specifically, the results of all methods, except the CE equating method results for the scores with low frequencies, are close to each other, and are around the DTM of 0.5. And, the RSD_R results of all methods (Case 11 – in Figure 32), again except the CE equating method results for the scores with low frequencies, are fairly close to each other, and are well below the DTM of 0.5. For ES=0.1 condition with DIF in both forms and 0.95 correlation (Case 14 – in Figure 33), the RSD_F means of all methods, again except the results of the CE equating method for the score ranges with low frequencies, are close to each other, and are around the DTM of 0.5. Also, the RSD_R results of all methods (Case 14 – in Figure 34), except the results of the CE equating method for the score ranges with low frequencies, are very close to each other, and are well below the DTM of 0.5. When ES changes to 0.3 (Case 17 – in Figure 35), the RSD_F results are almost the same. That is, the results of all methods, except the CE equating method results for the scores with low frequencies, are close to each other, and are around the DTM of 0.5. Also, the RSD_R results of all methods (Case 17 – in Figure 36), again except the CE equating method results for the scores with low frequencies, are very close to each other, and are well below the DTM of 0.5. To sum up, the RSD results of the methods are quite similar for 0.1 and 0.3 effect size conditions with DIF in both forms and various correlations.

For DIF in new form only, and various correlation conditions, the effects of group mean ability differences on equating invariance are as follows. Based on the RSD_F results, for ES=0.1 condition with DIF in new form only and 0.5 correlation

(Case 3 – in Figure 25), the SMO, and the CE equating method for high frequencies behave similarly to the criterion equating relationship. Also, for the RSD_R results (Case 3 – in Figure 26), the SMO method gives the closest results to the criterion equating relationship. When ES changes to 0.3 (Case 6 – in Figure 27), the RSD results almost remain the same. That is, for the RSD_F results, the SMO method, and the CE equating method for the scores with high frequencies behave similarly to the criteria. For the RSD_R results (Case 6 – in Figure 28), the method that gives the closest results to the criteria is the SMO. For ES=0.1 condition with DIF in new form only and 0.8 correlation (Case 9 – in Figure 29), IRT methods' RSD_F results are very close to each other, and are around the DTM of 1. The CE equating method results are also around the DTM of 1, but this is only valid for the scores with high frequencies. On the other hand, the RSD_R results of all methods (Case 9 – in Figure 30), except the CE equating method results at the scores with low frequencies, are similar, and are below the DMT of 0.5. When ES changes to 0.3 (Case 12 – in Figure 31), the results again remain the same. To mention in detail, IRT methods' RSD_F results are very close to each other, and are around the DTM of 1. Also, the CE equating method results are around the DTM of 1, but this is only valid for the scores with high frequencies. Additionally, based on the RSD_R results (Case 12 – in Figure 32), it is very difficult to distinguish the methods. That is, all methods' results, except the CE equating method results for the scores with low frequencies, are similar to each other, and are below the DTM of 0.5. For ES=0.1 condition with DIF in new form only and 0.95 correlation (Case 15 – in Figure 33), IRT methods' RSD_F results are very close to each other, and are around the DTM of 1. Additionally, the CE equating method results, for the scores with high frequencies, are also close to the DTM of 1. In other words, all methods, except the CE equating method for the scores with low frequencies, behave similarly for this case. Also, for the RSD_R results (Case 15 – in Figure 34), all methods, except the CE equating method for the scores with low frequencies, behave similarly, and give close results to each other. When ES changes to 0.3 (Case 18 – in Figure 35), the RSD_F and RSD_R results remain the same. Specifically, for RSD_F results, all methods behave similarly, and the results of all methods, except again the CE equating method results for the scores with low frequencies, are around the DTM of 1. On the other hand, for the RSD_R results (Case 18 – in Figure 36), all methods, except the CE equating method for the scores with low frequencies, behave similarly, and give close results to each other.

According to the findings, the group-to-overall conditional equating invariance (the RSD_F, and the RSD_R) results for all conditions with group mean ability difference between two forms show the same patterns as the RMSD results. Specifically, group mean ability differences between two forms do not have an effect on the relationship between DIF and equating invariance of the methods. The probable reason for this is that the differences in group abilities between forms are the same for total, reference and focal groups. On the other hand, for the RSD results, the magnitudes of equating dependence are larger for the focal group as compared to the reference group, and also the fluctuations of equating dependence across the scores are more dramatic for again the focal group.

Summary of Findings

This study is a simulation study which includes various simulation conditions including differential form DIF, correlation between dimensions, and group mean ability differences between two forms. CINEG design is used in this study. Data are generated according to SS-MIRT models. Equating procedures are conducted by using SMO, UO, UT, and EQ (CE equating with log-linear pre-smoothing) methods. And then, equating methods are compared with respect to their equating invariance results and the criterion equating relationship results. Here, the results of the SMO method, that is conducted with true parameters (generated parameters - not estimates), are used as the criterion equating relationship. Finally, results obtained from various conditions are interpreted in detail in the Findings section. Also, a brief summary of the findings is given below.

According to the findings, for 0.5 correlation condition, the method which demonstrates the effect of DIF on equating invariance most accurately is the SMO. It can be said that, the results of the CE equating method are also quite good and even sometimes the CE equating method is superior to the SMO. However, this is just valid for the scores with high frequencies. On the other hand, 0.8 correlation is the cut point of this study. That is, for 0.8 correlation and above, all methods behave similarly to each other, and give close results to the criterion equating relationship. Besides, when DIF added to both forms, the RMSD, the RSD_R and the RSD_F results are almost the same as those for no-DIF condition. But, adding DIF to one form only, increases the RMSDs, and group-to-overall indices. Specifically, the RMSD results

reach around the DTM of 0.5 while the RSD_F results almost reach around the DTM of 1. And, the RSD_R results approach the DTM of 0.5. On the other hand, group mean ability differences between two forms do not have an effect on the relationship between DIF and equating invariance. That is, for ES=0.3 condition the RMSD, the RSD_R and the RSD_F results are almost the same as those for ES=0.1 condition.

Chapter 5

Conclusion, Discussion, and Suggestions

This study examines the relationship between DIF and equating in multidimensional perspective. To reveal this relationship most accurately, equating invariance results were investigated for various DIF conditions. A major reason to examine equating invariance is to gain a better understanding of what happens when we use various equating methods with multidimensional tests which include DIF in common items. As equating procedures, MIRT, UIRT and EQ equating methods were used. And, the results of SMO equating with true parameters, were used as the criterion equating relationship of this study. Then, equating invariance results of MIRT, UIRT, and EQ equating methods were compared both within each other and with the criterion equating relationship. Finally, according to the findings, the methods which express the relationship between DIF and equating, most accurately were detected. And, for the fairness of test results, researchers and practitioners are advised to use this/these method(s) under certain conditions. From this point of view, the results of this study are discussed with the results of previous studies, and suggestions of the current study are presented below.

For the first research question, it was aimed to compare MIRT, UIRT, and EQ equating methods with respect to the effect of differential form DIF on population invariance of these methods. This condition was formed in this study to investigate the effect of difficulty difference between forms on population invariance, in MIRT perspective. According to the findings, when DIF is added in the same amount and in the same direction to the same common items in both forms, equating invariance of the methods are not affected. That is, the equating invariance results are similar to those for no-DIF condition. However, when DIF is added to one form only, the equating dependence of the methods increases and reaches the critical value. And, this may result in possible problems in validity and fairness of the reported scores. It has been shown in previous studies that differential difficulty across test forms may result in equating dependence in CTT (Cook & Petersen, 1987; Dorans, 2004), and in IRT (Huggins, 2012, 2014). The results of the current study support these findings. Specifically, Cook and Petersen (1987) emphasized that the difficulty difference in common items of the new and old forms increases the equating dependence. Therefore, they stated that anchor items should be examined in terms

of difficulty differences between two forms. Likewise, the research results of Huggins (2012, 2014) show that the difficulty difference in anchor items for different forms affects the equating invariance of equating methods based on IRT. While Huggins' study emphasizes the relationship between DIF and unidimensional IRT-based equating methods, current research goes one step further and examines this relationship in terms of multidimensional IRT. Also, Dorans (2004) discussed in detail the relationship between DIF and equating invariance. According to this study, differential difficulty across two forms may cause lack of equating invariance. However, this research includes theoretical information only. On the other hand, in the current research this theoretical knowledge has been proven on different applications. Dorans also emphasized that DIF and equating invariance studies should be carried out together for fair assessments. A similar result is presented in the suggestions of current research. Besides the mentioned studies, Dorans, Lin, Wang and Yao (2014) examined the effects of multidimensionality on latent score and observed score linking results. According to the results, if the two tests had parallel structure (no difference in test difficulty and content structure between two tests), then the linking relationship between two forms was invariant across different sub-populations. However, when the parallel structure disappeared, sub-population invariance was not achieved, especially under the 0.5 correlation condition. These results are quite consistent with the results of the present study. In other words, the difficulty differences between two forms cause equating dependence.

On the other hand, Atalay Kabasakal and Kelecioglu (2015) emphasized that equating error (RMSE) and equating bias (BIAS) of the equating methods (MIRM, IRM-CC, and IRM-SC) increased in the presence of high-level of DIF. Also, Demirus and Gelbal (2016) used DIF and linking together in their study. According to the results of this research, when DIF was added to the anchor items, equating errors of the mean-mean, Haebara, and Stocking-Lord methods increased. Also, Yurtcu and Guzeller (2018) investigated the equating errors of mean-mean, mean-sigma, Haebara, and Stocking-Lord linking methods in the presence of DIF. According to the results of this study, equating errors increased when DIF was added to one form only. Although the results of mentioned studies seem similar to the results of the current study, there are fundamental differences between these studies. One of these is the evaluation criteria used in the aforementioned studies. These criteria

represent equating errors. However, the current study examines population invariance in the presence of DIF to reveal the relationship between DIF and equating. Hence, equating invariance indices were used in the current study as evaluation criteria. In this perspective, the results expressed by the current study are different from mentioned studies.

Second research question is formed to compare MIRT, UIRT and EQ equating methods with respect to the effect of correlation between dimensions on population invariance of these methods. With 0.5 correlation condition, the method that gives the closest results to the criterion equating relationship among the compared equating methods is the SMO. UO and UT give high equating invariance indices results even in the conditions with no-DIF. These results are consistent with Ackerman's (1992) study. Ackerman (1992) stated in his study that using UIRT methods with multidimensional data sets and treating test scores as if they reflect a single dimension may cause item bias. This is the reason why the population invariance values of UIRT methods in the present study are much higher than MIRT method. On the other hand, CE equating behaves very close to the criterion at the scores with high frequencies, but deviates far from the criterion at the scores with low frequencies. Here, in this study, the scores with low frequencies correspond to very low (close to 0) and very high scores (close to 80). In 0.8 and 0.95 correlations, the results of the methods are very close to each other. In summary, it can be said that for 0.5 correlation the method that shows the effect of DIF on the equating invariance most accurately (closest to the criterion) is the SMO, and for correlations of 0.8 and above, all methods reflect this effect appropriately. The results for the correlation levels are consistent with previous studies that were conducted by using various MIRT equating methods. Specifically, Lee and Brossman (2012) emphasized in their study that when data were multidimensional, the SMO method produced adequate equating results and outperformed UIRT procedure. Also, Lee, Lee and Brennan (2014) specified that MIRT procedure provided more accurate equating results than other equating procedures especially when the correlation between dimensions was low. Another study that used MIRT and UIRT equating procedures together belongs to Lee and Lee (2016). According to the results of this study, the MIRT equating method produced more accurate equating results than the UIRT equating method when a certain degree of multidimensionality existed (e.g.,

for 0.5 correlation). Kumlu (2019) also used UIRT and MIRT equating procedures in his study. According to the results, the MIRT equating method produced more accurate equating results than the UIRT equating method when the data was multidimensional. Kim, Lee and Kolen (2020) also stated in their study that MIRT equating methods showed more accurate equating results compared with the UIRT equating when the data were multidimensional. These studies mentioned, presented similar results with the current study. Additionally, according to Peterson and Lee (2014) the multidimensional equating methods were found to perform better for datasets that evidenced more multidimensionality. The result of this study so far is consistent with the results of the present study. However, Peterson and Lee (2014) also highlighted in the rest of their results that unidimensional methods worked better for unidimensional datasets. But, in the current study, in a correlation of 0.8 and above both multidimensional and unidimensional methods yield similar results, and neither one is superior to the other. It can be said that the results of the two studies differ from each other in this respect.

The above comparisons have focused on MIRT and UIRT methods. On the other hand, the comparison of IRT and non-IRT methods is as follows. For the present study, it can be said that non-IRT method (CE equating method) give close results to IRT methods and even sometimes better results than IRT methods. Specifically, when the correlation between dimensions is 0.5, CE equating method gives closer results to the SMO method than the UIRT method. However, this case is valid at the scores with high frequencies. The reason why CE equating method results differ according to frequency is that it does not use parameter estimates as IRT methods do, on the contrary, it conducts equating by using frequencies based on the number correct scores. Hence, considering the whole score range, IRT methods always give better results. Lee and Brossman (2012) state that the results of SMO procedure were more similar to the results of equipercentile procedure than to the results of UIRT procedure. According to them one plausible explanation for this observation might be that when the data were not strictly unidimensional, the UIRT procedure violated the assumption of unidimensionality. On the other hand, the SMO procedure took into account the multidimensionality, and the equipercentile procedure did not have an assumption about dimensionality, hence was less effected by the dimensional structure of the test. Considering the entire

score range for the current study, it can be said that the results of these two studies are different from each other. On the other hand, Lee, Lee and Brennan (2014) stated in their study that the IRT-based equating procedure seemed to perform better than the EQ equating procedure. Peterson and Lee (2014) also stated that the MIRT and UIRT methods behaved close to each other, but the EQ method behaved slightly different from them. In this respect, the results of the two studies mentioned are consistent with the results of the present study.

Finally, it should also be noted that, comparison results for different equating methods always depends on what criterion to use. In this study, the criterion is based on IRT, and hence it is in favor of IRT methods. This important point should always be taken into account in the comparison studies.

Third research question is added to this study with the aim of comparing MIRT, UIRT, and EQ equating methods with respect to the effect of group mean ability differences between two forms on the relationship between DIF and population invariance of MIRT, UIRT and EQ equating methods. For the focal, reference and total group, the group mean ability difference (ES) between two forms are formed as 0.1 and 0.3. For example, in the first condition group abilities are formed as $(\theta_1, \theta_2)_{Old}$ ~ *BN*(0, 0, 1, 1, *ρ*), $(\theta_1, \theta_2)_{Old_F}$ ~ *BN*(-.3, -.3, .9, .9, *ρ*), $(\theta_1, \theta_2)_{Old_R}$ ~ *BN*(.1, .1, .99, .99, *ρ*) in the old form, and for ES=0.1 condition they changed to $(\theta_1, \theta_2)_{New}$ ~ *BN*(.1, .1, 1, 1, *ρ*), $(\theta_1, \theta_2)_{New_F}$ ~ *BN*(-.2, -.2, .9, .9, *ρ*), $(\theta_1, \theta_2)_{New_R}$ ~ *BN*(.2, .2, .99, .99, *ρ*) in the new form. According to the findings, both ES=0.1, and ES=0.3 conditions do not have an effect on the equating invariance results of the methods. One plausible explanation for this observation might be that when the difference in group abilities between old and new form is equal in focal, reference, and total group, the equating invariance of the methods do not change. That is, these equal differences in the sub groups and in the total group do not make a difference on the equating invariance results of the methods. These findings are compared with other research findings in the literature. First of all, Lee et al. (2012) used EQ equating methods in their study. With respect to the results, they stated that the error increased as the effect size of the group mean ability differences increased. In this respect, the research results are different from the current research results. According to Peterson (2014), (M)IRT methods were robust to large group mean ability differences while the traditional equipercentile method was

affected from this. However, the following point was emphasized in this research that there were mixed findings in the literature on whether traditional or IRT equating methods were more robust to group differences. So, it was difficult to hypothesize the extent to which the presence of group differences affected each of the studied equating methods in that research. And, it was difficult to make that judgment since the population equating relationship was unknown. On the other hand, the results of the current research are as follows: both effect size of 0.1, and 0.3 do not have an effect on the equating invariance results of the (M)IRT and EQ equating methods. And these results are based on the population equating relationship that Peterson mentioned in her study. In this respect, the results of the current study will help to eliminate the confusion in the literature. In summary, the results of the third research question are also very important and useful for the psychometric literature. Finally, Kim, Lee and Kolen (2020) also used effect size of 0.1 and 0.3 in their study. And, according to the results, SS-MIRT procedures (SMO and SMT) seemed relatively robust to large group difference. In terms of the SMO method, the results of this study and the results of the current study support each other.

Above, the findings of the current study are summarized and compared with the findings of other similar studies. Based on the conclusions mentioned, various methods and their performances in the presence of DIF and multidimensionality conditions are discussed. To ensure score equity of the reported scores, this study helps practitioners to be careful with some specific conditions. First, it is advisable to monitor for anchor items that display differential form DIF, because using these items in reporting equating relationships may lead to problematic levels of equating dependence. Second, many tests are multidimensional, and there are many sources of DIF. The structure of the test should be investigated first, and if it turns out to be multidimensional, it is recommended to use multidimensional equating methods because they are thought to give the most reliable results in equating invariance studies. And third, the population invariance results of both the SMO and EQ equating methods do not change according to 0.1 or 0.3 effect size conditions for both sub groups and the total group. However, more studies are needed on this subject to confirm these results.

Suggestions

The results of this research are very important for both practitioners who will equate multidimensional tests and researchers who will work on this subject. In this respect, various suggestions are presented to researchers and test practitioners in this part of the study. Suggestions based on the research results, and suggestions for future research are presented below.

Suggestions based on the research results

1. As mentioned above, it is recommended to use multidimensional equating methods in the presence of multidimensional data. Thus, whether the equating results are group dependent or not under various DIF conditions caused by multidimensionality can be reflected most accurately. At the correlation conditions of 0.8 and above, any of the (M)IRT methods can be used, as each of these methods properly reflect the effect of DIF on equating. However, this is not valid for the EQ method. Because, this method gives very high population invariance results at the scores with low frequencies. Therefore, the EQ method is only suggested to be used at the score ranges with high frequencies.

2. Both, DIF in both forms (same direction and same amount) and no-DIF conditions create the same effect on the population invariance of the equating results. On the contrary, differential form DIF may cause equating dependence. In this respect, especially for differential form DIF conditions, the equating results should be monitored in terms of population invariance. And, as Dorans (2004) stated, to ensure the fairness of the equating results, DIF and population invariance studies should be conducted together.

3. (M)IRT or EQ methods can be used when the group mean ability difference between the two forms is 0.1 or 0.3 for the subgroups and total group. Because, the results of these methods are not affected by group mean ability differences. However, in this study, group differences between two forms are the same for the subgroups and total group. In this respect, more research is needed on this issue.

Suggestions for future research

1. For the MIRT model, in some iterations the number of cycles was not sufficient and the convergence criterion could not be met, hence, these replications

were repeated. On this issue, it can be suggested to increase the MaxE (maximum allowed number of E-steps in the EM algorithm) and MaxM (the number of allowable iterations in each of the iterative M-steps) values in the flexMIRT input in order to obtain more successful estimations.

2. The time which is required for SS-MIRT estimations is another issue of this study. Also, conducting concurrent calibration for the CINEG design requires additional time. Hence, it is recommended that researches who will use MIRT models should be aware of the time limitation.

3. Another important issue of this study is the equating criterion. Unfortunately, a perfect equating criterion does not exist in the literature. In this study, SMO method with true parameters was used to reflect the criterion equating relationship. And, the DTM of 0.5 and 0.1 were used to reflect the degree of equating dependence of the methods. Employing SMO as the equating criterion of this study is thought to be in favor of the IRT methods. It is also seen from the results that EQ equating method behaved differently from the other methods at the scores with low frequencies. Because, EQ equating method conducted equating by using the frequencies based on the number correct scores while IRT methods used parameter estimates. From this point of view, the comparison between non-IRT and IRT methods should be done with caution.

4. Besides, future studies may be conducted by using various levels of sample size, magnitude of DIF, and numbers of DIF items.

5. It is also recommended to examine the equating results using the bi-directional DIF. Thus, the DTF effect created by bi-directional DIF items and its effect on the equating results can be discussed.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Atalay Kabasakal, K., & Kelecioğlu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory & Practice, 15*(5), 1229-1246.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Addison-Wesley.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of the EM algorithm. *Psychometrika, 46*, 443-459.

Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement, 32*, 102-114.

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*, 460-481.

Cai, L. (2017). *flexMIRT* (Version 3.51). Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications, Inc.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20*, 405–416.

Demirus, K. B., & Gelbal, S. (2016). Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(1),182-201.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Dorans, N. J. (2008, March). *Three facets of fairness.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Dorans, N. J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81-97.

Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (ETS Research Report No. RR-14-41). Educational Testing Service.

Dorans, N. J., Moses, T. P., & Eignor D. R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Educational Testing Service.

Dunlap, J. W. (1937). Combinative properties of correlation coefficients. *The Journal of Experimental Education, 5*(3), 286-288.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). American Council on Education.

Gibbons, R. D., Bock, R. D., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423-436.

Green, B. F. (2003). Comments on population invariance of score linking. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR-03-27, pp. 127-130). Educational Testing Service.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note. Iowa City, IA: ACT, Inc.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*, 244-253.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Lawrence Erlbaum.

Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational Measurement: Issues and Practice, 31*(1), 27-40.

Huggins, A. C. (2012). *The effect of differential item functioning on population invariance of item response theory true score equating* (Publication No. 3511759) [Doctoral dissertation, University of Miami]. ProQuest Dissertations & Theses Global.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement, 74*(4), 627-658.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Kim, S. Y. (2018). *Simple structure MIRT equating for multidimensional tests* [Doctoral dissertation, University of Iowa]. University of Iowa's Institutional Repository. https://ir.uiowa.edu/etd/6158/

Kim, S. Y., Lee, W. C., & Kolen, M. J. (2020). Simple-structure multidimensional item response theory equating for multidimensional tests. *Educational and Psychological Measurement, 80*(1), 91-125.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3-14.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd edition). Springer.

Kolen, M. J., Wang, T., & Lee, W. C. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*, 1-20.

Kumlu, G. (2019). *Test ve alt testlerde eşitlemenin farklı koşullar açısından incelenmesi* [Doctoral dissertation, Hacettepe University]. Hacettepe University Institutional Repository. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/8877

Lee, W. C., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen & W. C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 2)* (CASMA Monograph No. 2.2.) Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Lee, G., & Lee, W. C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29*, 224-241.

Lee, Y. H., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing, 17*(1), 23-54.

Lee, G., Lee, W. C., Kolen, M. J., Park, I. Y., Kim, D. I., & Yang, J. S. (2015). Bi-factor MIRT true-score equating for testlet-based tests. *Journal of Educational Evaluation, 28*, 681-700.

Lee, W. C., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 2)* (CASMA Monograph No. 2.2.) Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Lee, E., Lee, W. C., & Brennan, R. L. (2014). *Equating multidimensional tests under a random groups design: A comparison of various equating procedures.* (CASMA Research Report No. 40). Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*, 3-25.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115-138.

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*(4), 647-677.

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of k–12 large-scale science assessment. *Journal of Applied Testing Technology, 13*, 1-27.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement, 8*, 452-461.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147–176). Academic.

Mazor, K. M., Hambleton, R. K., and Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357–367.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82-1). American College Testing.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Erlbaum.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: four practical approaches. *Journal of Educational Measurement, 37*, 357-373.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Petersen, N.S. (2008). A discussion of population invariance of equating. *Applied Psychological Measurement, 32*, 98-101.

Peterson, J. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests* [Doctoral dissertation, University of Iowa]. University of Iowa's Institutional Repository. https://ir.uiowa.edu/cgi/viewcontent.cgi?article=5418&context=etd

Peterson, J., & Lee, W. C. (2014). Multidimensional item response theory observed score equating methods for mixed-format tests. In M. J. Kolen & W. C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Volume 2) (CASMA Monograph No. 2.3). Center for Advanced

Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Powers, S., & Kolen, M. J. (2014). Evaluating equating accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement, 51*(1), 39-56.

R Core Team (2016). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics.* (pp. 607–642). Elsevier.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dtf as well as item bias/dif. *Psychometrika, 58*(2), 159-194.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, Minneapolis.

Tao, W., & Cao, Y. (2016). An Extension of IRT-Based Equating to the Dichotomous Testlet Response Theory Model. *Applied Measurement in Education, 29*(2), 108-121.

Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and poststratification methods for observed-score equating and their relationship to population invariance. *Journal of Educational Measurement, 41*- 15-32.

von Davier, A. A., & Liu, M. (2006). *Population Invariance of Test Equating and Linking: Theory Extension and Applications Across Exams* (ETS Research Report No. RR-06-31). Educational Testing Service.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11-26.

Yang, W. L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33-41.

Yao, L. (2011). Multidimensional Linking for Domain Scores and Overall Scores for Nonequivalent Groups. *Applied Psychological Measurement, 35*(1), 48-66.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 1–23.

Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*, 177-197.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*, 469-492.

Yurtçu, M., & Güzeller, C. O. (2018). Investigation of equating error in tests with differential item functioning. *International Journal of Assessment Tools in Education, 5*(1), 50-57.

Zhang, J. (2004). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (ETS Research Report 04-44). Educational Testing Services.

Zhang, O. (2012). *Observed score and true score equating for multidimensional item response theory under nonequivalent group anchor test design* [Doctoral dissertation, University of Florida]. University of Florida Institutional Repository. https://ufdc.ufl.edu/UFE0044500/00001/pdf