Id: 1031745

Participatory Educational Research (PER)
Vol.9(5), pp. 143-160, September 2022
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.22.108.9.5

Effect of Item Parameter Drift in Mixed Format Common Items on Test Equating

İbrahim Uysal[*]
Educational Measurement and Evaluation, Bolu Abant İzzet Baysal University, Bolu, Turkey
ORCID: 0000-0002-6767-0362

Merve Şahin-Kürşad
Educational Measurement and Evaluation, National Defence University, Ankara, Turkey
ORCID: 0000-0002-6591-0705

Abdullah Faruk Kılıç
Educational Measurement and Evaluation, Adıyaman University, Adıyaman, Turkey
ORCID: 0000-0003-3129-1763

Article history
Received:
02.11.2021

Received in revised form:
05.01.2022

Accepted:
20.05.2022

Key words:
Mixed format tests,
Item parameter drift,
Test equating,
Mixed format common test,
Monte Carlo simulation,
CINEG design

The aim of the study was to examine the common items in the mixed format (e.g., multiple-choices and essay items) contain parameter drifts in the test equating processes performed with the common item non-equivalent groups design. In this study, which was carried out using Monte Carlo simulation with a fully crossed design, the factors of test length (30 and 50), sample size (1000 and 3000), common item ratio (30 and 40%), ratio of items with item parameter drift (IPD) in common items (20 and 30%), location of common items in tests (at the beginning, randomly distributed, and at the end) and IPD size in multiple-choice items (low [0.2] and high [1.0]) were studied. Four test forms were created, and two test forms do not contain parameter drifts. After the parameter drift was performed on the first of the other two test forms, the parameter drift was again performed on the second test form. Test equating results were compared using the root mean squared error (RMSE) value. As a result of the research, ratio of items with IPD in common items, IPD size in multiple-choice items, common item ratio, sample size and test length on equating errors were found to be significant.

Introduction

Tests, in education, have various areas of use, including selecting students for specific programs, monitoring students' development, and identifying their status. Large-scale tests that are periodically implemented to monitor students' development and compare students may benefit from parallel test forms to ensure confidentiality and reduce cheating attempts (Chen,

[*] Correspondency: ibrahimuysal06@gmail.com

2013). These forms should yield valid and reliable results, and the attained scores should be comparable (Kilmen, 2010). For the test forms to be comparable on the same scale, two important processes of psychometrics are required, namely scaling and equating (Guo, Zheng & Chang, 2015).

Test equating is an important process that ensures the validity of different test forms. Many international testing programs such as Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Trends International Mathematics and Science Study (TIMSS) uses test equating to monitor student achievement over the years. Like in these large-scale tests, many testing conditions should have high level of security. For this purpose, different forms of tests can be created and then compared by test equating process. Two tests that measure the same construct using test equating that is defined as converting one-unit system of a test form to another test form's unit system (Angoff, 1971) are correlated with each other (Felan, 2002).

Test equating based on item response theory (IRT) consists of three significant components, namely item parameter estimation, rescaling of item parameters, and scale conversion (Kolen & Brennan, 2004). One requirement of the IRT-based test equating is to study the item parameter invariance assumption that examines the item parameter invariance in multiple tests taken by the same group of students (Gaertner & Briggs, 2009). However, the students may fail to obtain item parameter invariance assumption due to reasons such as incorrect item calibration, miscalculation of item location on the scale, overuse of items, cheating, security flaws, etc. (Li, 2008; Stahl & Muckle, 2007). This situation is defined as item parameter drift (IPD) in literature.

IPD causes errors in the measurement results since it violates the invariance assumption, which is one of the main assumptions of IRT. Validity decreases when variables that are irrelevant to the construct, which is to be measured, are included in measurement results (McCoy, 2009). This leads to certain problems in the administration of tests that require the invariance property in item parameters, including test developing and test equating. In administering test equations, IPD may cause an increase in measurement error by leading to errors in item calibration (Li, 2008; Meng, Steinkamp & Matthews-Lopez, 2010).

While test equating uses various methods such as random groups and single group design, this study focuses on common item non-equivalent groups design (CINEG) method for anchor tests, because many international testing programs such as PISA, PIRLS, TIMSS use mixed format common tests. This design equates two test forms over common items that are used to establish the relationship between different test forms that the two groups take. Common items may have a mixed format (multiple choice and open-ended) in CINEG design. It is important for these common items to represent test forms statistically and content-wise. The most significant assumption of the CINEG design is the maintenance of item parameters of common items that are used in equating, and remains constant between forms (Kolen & Brennan, 2004; Michaelides, 2010). However, item parameter drift may occur due to estimation error, context effect, exposure of items, and overuse of items that lead to differentiation in the parameters of common items (Li, 2012).

IPD occurrences in test equating studies yield two significant results both of which are due to its effects on the item parameter estimations. One of them is biased ability estimations since IPD's effect on item parameter estimations disrupt parameter invariance assumption. Another result is the effect IPD has on the equating coefficient by affecting item parameter estimations.

It is, thus, significant to identify the IPD conditions that negatively affect test equating (Li, 2012) to avoid erroneous test equating. While some studies in literature examine the impact of IPD on test equating (Arce-Ferrer & Bulut, 2017; Chen, 2013; Demirus & Uysal, 2016; Gaertner & Brigss, 2009; Han, Wells & Hambleton, 2015; Michaelides, 2010), most studies focus on the results of IPD occurrences in -a or -b parameters in multiple choice items (Demirus & Uysal, 2016; Han & Guo, 2011; Han & Wells, 2007; Wells, Subkoviak & Serlin, 2002). On the other hand, there are very few studies that examine the impact of IPD on test equating, particularly when common items are in a mixed format (Hu, Rogers & Vukmirovic, 2008; Li, 2012).

Bock, Muraki and Pfeiffenberger (1988) conducted a study on the maintenance of IRT scale in the presence of IPD and they found that IPD came up in time especially in large sample sizes. Wollack, Sung and Kang (2005) examined the longitudinal impact of IPD in a German placement test where scaling was carried out between administration of tests. They emphasized that the scaling method and IPD identification model had significant effects on the individuals' ability estimations and capacity to pass decisions. They supported their results with another study that they carried out a year later in which Wollack, Sung and Kang (2006) found that direct scaling yielded better results than indirect scaling in the presence of IPD. They underlined that test characteristic curve and scaling methods performed better than fixed parameter linking methods. Han (2008) conducted a study related with IPD was seen in the equating process. With series of simulations, Han (2008) found that when most of the common items (50%) had high level of IPD (1.00 logit), ability estimations were impacted substantially. Similarly, Lee and Geisinger (2019) analyzed the effects of IPD on three different lengths of TIMSS 2011 scales and they determined that scales with fewer items have more impact on ability estimation and the improper classification of individuals.

Chen (2013), who compared the performance of equating methods and used the three-parameter logistic model (3PLM) as well as non-equivalent groups with anchor tests (NEAT) function to keep or exclude items in and from tests in the presence of IPD, stated that IPD type had very small effects on the fixed common item parameter (FCIP) method. The impact of IPD on the Stocking-Lord test characteristic curve method (TCC-SL) and the concurrent method varied depending on whether the item that displayed IPD were kept in or excluded from the test. Accordingly, for the sake of correct estimations, when using the concurrent method, excluding the items that displayed IPD from scaling yielded better results. Demirus and Uysal (2016) compared test equating methods (mean-mean, mean-sigma, SL, Haebara) with a high level of parameter drift in the -b parameter in their simulation study with 3PLM. The b parameter created one-directional increasing, one-directional decreasing, and multidirectional (both increasing and decreasing) parameter drifts. The SL method showed the lowest test equating error (RMSD) when one-directional increasing and multidirectional parameter drift were observed. The mean-mean method showed the lowest test equating error with one-directional decreasing parameter drift.

Han et al. (2015) examined the impacts of IPD on test equating in parameter -c of common items and observed a negative effect on equating results when 32 different conditions were compared using Monte Carlo simulation methods. When using the 3PLM in particular, the occurrence of IPD in parameter -b of common items affected the estimations of both -a and -c parameters. However, the full impact of IPD was not observed on parameter -b when IPD was uniform since a part of IPD was absorbed by parameter -c. In the studies so far, different conditions, such as ratio of common items, test length, sample size etc., were studied in IPD scope. However, these studies have been conducted in the extent of multiple-choice items.

Several studies on test equating and IPD separately examined the conditions where common items consisted of multiple-choice items as well as items of mixed formats. Hu et al. (2008) examined the impact of parameter drift in parameter -b of common items by using IRT-based test equating methods on test equating and the performances of equating methods when parameter -b comprised extreme values that displayed too much parameter drift. Their simulation study adopted the CINEG design, and group ability distribution, score and number of extreme values, type of extreme values as well as variables of equating methods were manipulated. Common items comprised 10 items, out of which 8 were multiple choice, 1 was a short-answer, and 1 was an open-ended item. A test form comprised a total of 36 items. Data were simulated to allow for two different test forms to be administered one year apart. The study results showed that methods of concurrent calibration, separate calibration with test characteristic curve (TCC), and calibration with fixed common item parameters (FCIP) yielded fewer errors compared to conditions that ignored IPD except for the mean-sigma method where IPD was observed, and extreme values were weighted when IPD occurred in parameter -b. Therefore, evidently, equating by observing IPD occurrence resulted in fewer equating errors.

Li (2012) examined the impact of IPD on test equating when common items included mixed format items. The simulation study generated data based on the conditions of large-scale administration of tests. Factors such as IPD size, number of common items, number of open-ended items with IPD, and ability distribution of former and latter groups who took the test were manipulated. The study found that among the four listed factors, while ability distribution of the group had the least impact on scaling and equating, numbers of common and open-ended items with IPD were the factors with the most impact. Scaling coefficients yielded more erroneous results when two open-ended items displayed IPD rather than one, proving that open-ended item with IPD should be excluded from the test altogether. However, an increased number of common items led open-ended items with IPD to have a lower effect on error values.

Although studies in literature revealed that IPD led to errors in test equating, a large number of these studies focused on conditions where the common items consisted of multiple-choice items. There were only few studies that focused on conditions where common items had a mixed format, and these studies examined the impact of IPD on equating for a limited number of conditions. Aside from factors such as item parameter drift ratio in common items, sample size, test length, and item parameter drift size, this study—as distinct from other studies—examined the impacts of IPD by accounting for the location of common items in the tests. The impacts of IPD on test equating errors were identified by using ability estimations. To this end, this study sought answers to the research questions below:

(1) What are the impacts of IPD size and ratio of common items in equating errors for equating processes that use common item non-equivalent groups design in anchor tests?
(2) What are the impacts of test length, sample size, ratio of common items, and location of common items in tests on test equating errors in tests with IPD?

Method

Research Design

This is a Monte Carlo simulation study that follows an experimental process, generates repeated random samples, inspects all conditions that are not examined in the study, and compares theoretical models. Monte Carlo simulation methods are carried out based on real life population parameters (Carsey & Harden, 2014; Mooney, 1997).

Simulation Conditions

The controlled simulation conditions of this study, which are shown in Figure 1, are ability distribution (θ~N [0,1]); ability estimation method (expected a posteriori [EAP]); type of test (mixed format tests featuring both dichotomous and polytomous items); item response theory models (3PLM for dichotomous items and generalized partial credit model [GPCM] for polytomous items); IPD sizes for polytomous items (see data generation); test equating method (Stocking-Lord [SL]); and equating design (common item non-equivalent group design [CINEG]).

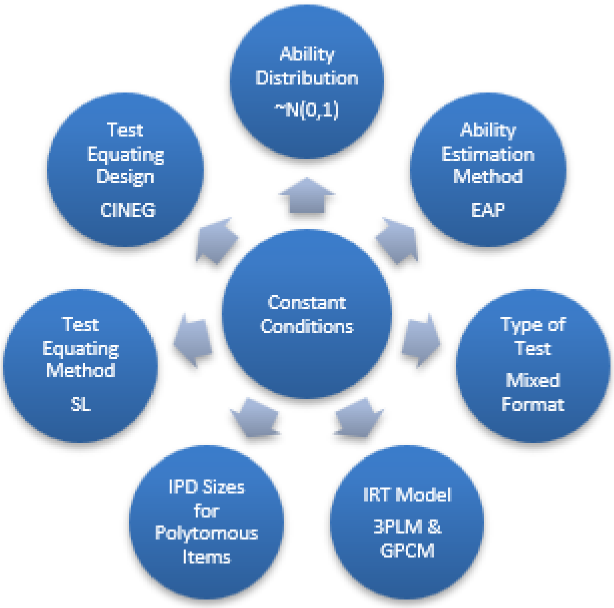
Ability Distribution ~N(0,1)
Ability Estimation Method EAP
Type of Test Mixed Format
IRT Model 3PLM & GPCM
IPD Sizes for Polytomous Items
Test Equating Method SL
Test Equating Design CINEG
Constant Conditions

Figure 1. Controlled simulation conditions.

In the research, we used mixed format tests because open-ended and multiple-choice items eliminate each other's disadvantages (Messick, 1993). We used EAP for ability estimation. The EAP method is the most commonly used estimation method in unidimensional structures (Brown & Croudace, 2015). It is known that the EAP method makes more accurate estimations than the maximum likelihood (ML) method (Sass, Schmitt & Walker, 2008). In addition, the EAP method can make more accurate estimations than the MAP method in polytomous models (Kieftenbeld & Natesan, 2012). Sass et al. (2008) stated that the abilities should follow normal distribution in order to accurately estimate abilities in IRT. Therefore, the distribution of ability in the study was determined as N(0,1). Since it is frequently used for test security and practical reasons, the CINEG design was used in this study (Kolen & Brennan, 2004). This study opted Stocking-Lord (SL) as its equating method since the characteristics curve methods (Haebara and SL) that were based on separate calibration and used in the CINEG design yielded better performances than moment methods (Hanson & Beguin, 2002; Kim & Lee, 2006). Studies have shown that SL yielded more correct results compared to equating methods, including moment methods of mean-standard deviation and Kernel equating method (Meng, 2012).

This study examines the impacts of test length (30 and 50 items), sample size (1000 and 3000), common item ratio in tests (30% and 40%), ratio of items with IPD in common items (20% and 30%), location of common items in tests (at the beginning, randomly distributed, and at the end), and IPD size in multiple-choice items (low [0.2] and high [1.0]) on test equating by manipulating them. Figure 2 illustrates the manipulated simulation conditions used in the study.

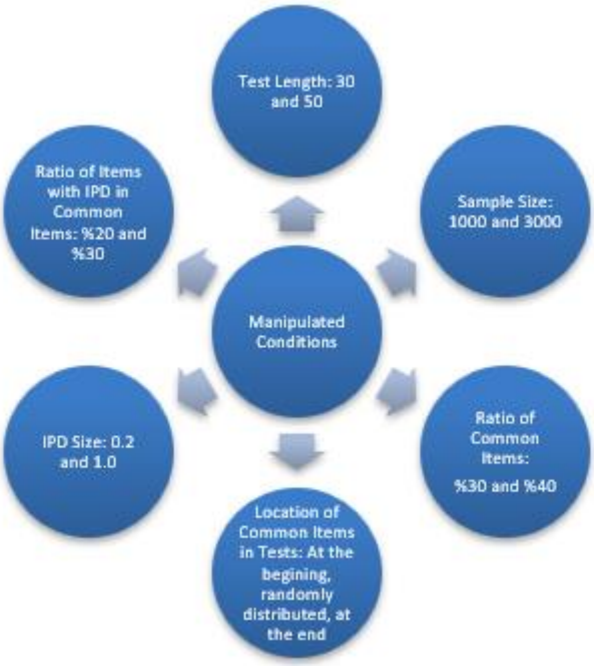
Test Length: 30 and 50
Sample Size: 1000 and 3000
Ratio of Common Items: %30 and %40
Location of Common Items in Tests: At the begining, randomly distributed, at the end
IPD Size: 0.2 and 1.0
Ratio of Items with IPD in Common Items: %20 and %30
Manipulated Conditions

Figure 2. Manipulated simulation conditions.

Kim, Walker and McHale (2010) studied 36 items in the test equating study they carried out on a large-scale test in mixed format. From this point of view, a condition of 30 items was added to the study. Then, we add the 50 items condition to the simulation because it is a typical test length used in standard tests such as the Florida Comprehensive Assessment Test (FCAT) (Tian, 2011). While determining the test length, it was considered that open-ended items gave more information. Since the study was conducted on a mixed-format test, the number of score points are as important as the number of items (Cao, 2008). In the current study, there are 48 number-correct scores in the condition where the test length is 30, and 80 number-correct scores in the condition that it is 50. Sample sizes were determined as 1000 and 3000 because three-parameter logistic model of IRT yielded consistent results with a minimum sample size of 1000 (Skaggs & Lissitz, 1986). Since studies recommended that the ratio of common items be 20% or more when the number of items in the test is 40 (Kolen & Brennan, 2004), the ratio of common items was set at 30 and 40%. The ratio of polytomous items to the number of items in tests was reported to be generally between 20 and 40% (Tian, 2011). Based on this, the ratio of polytomous items was identified as 20%.

IPD size in dichotomous items was determined as 0.2 (low) and 1.0 (high) in this study. It is difficult to detect when the IPD magnitude is below 0.5 (Han & Guo, 2011). However, in this study, an IPD size of 0.2 was added to the conditions since there was a different amount of IPD in open-ended items. In the literature, it was found that the IPD size was determined as 0.8 in the b parameter (Sukin, 2010). In the current study, the condition 1.0 was added to the study in order to examine the effect of higher level of IPD size together with IPD in open-ended items. In the current study, 20% and 30% of the common items have IPD, which means that approximately 2, 3, 4, and 6 of all items have IPD. Chen (2013) has studied cases where 10% and 25% of the common items have IPD. In the current study, rates were set below 10% because only common items have IPD. Although not an equating study performed in the same context, Marengo, Miceli, Rosato and Settanni (2018) stated that changing the location of the item with DIF in common items causes a difference in the difficulty of the items. Therefore, in the current study, in which internal common items were used, the positions of the items with parameter drift were changed. The study was carried out using fully crossed design with 96 conditions. Every condition was replicated 1000 times.

Data Generation

The study was carried out using four test forms where one of them was the baseline test form (Form 1). Of the other three, one did not contain IPD, and was generated as a reference point in test equating errors (Form 4). The remaining two test forms contained IPD, out of which Form 2 was re-exposed to IPD to create the second test form, Form 3.

Tests and common items had a mixed format that was recommended to equate processes in anchor tests in order to represent test characteristics and contents of common items (Kim & Lee, 2004; Tate, 2000). Additionally, the use of items with constructed responses as common items reduced equating errors (Tian, 2011). Parameter drift was generated in some of the common dichotomous and polytomous items. All manipulations in parameter drifts in the study were done by reducing. Similar to the current study, Sukin (2010) also reduced the a and b parameters for parameter drift. A parameter drift of -0.5 was generated for parameter -a of dichotomous and polytomous items. Similar to this research, Wells et al. (2002) used the same value for the drift in parameter a. In Form 2, drift was generated for the first two thresholds at -0.25 and for the last two thresholds at -0.10 in parameter -b of polytomous items. In Form 3, drift was at -0.25 for the first two thresholds and at -0.20 for the last two thresholds. In the current study, low level of parameter drift was performed at all thresholds in Form 2. In Form 3, on the other hand, since parameter drift were performed twice, a moderate parameter drift was observed in the first two thresholds, and a low-level parameter drift in the last two thresholds (Penfield, Alvarez & Lee, 2008).

Dichotomous items were generated using the 3PLM and polytomous items were generated using GPCM (Muraki, 1992). While dichotomous items (0, 1) were placed in the test with multiple-choice items in mind, polytomous items were placed with open-ended items rated in five categories (0, 1, 2, 3, and 4) in mind. Similar to this research, the items with constructed responses in the Advanced Placement Program consist of 5 categories (Wang, Drasgow & Liu, 2016).

In the literature, it was observed that 3PLM yielded more consistent results than 2PLM and 1PLM for dichotomous items and GPCM yielded more consistent results than PCM and GRM for polytomous items (Chon, Lee & Ansley, 2007). Moreover, 3PLM accounted for the correct answer being given by chance (DeMars, 2010). Anchor tests were thus composed using 3PLM and GPCM since GPCM was the most compatible model with 3PLM (Chon et al., 2007). When generating items with 3PLM distributions of $a{\sim}lnN$ (0.3, 0.2), b~N (0, 1) and $c{\sim}Beta$ (8, 32) were used. Bulut and Sunbul (2017) as well as Kim and Lee (2004) had also benefited from similar distributions for parameter values and distributions that were expected to represent large-scale tests. The discrimination parameter of polytomous items was the same as that of dichotomous items, and the location parameter was identified as d1~N (-1.5, 0.2), d2~N (-0.5, 0.2), d3~N (0.5, 0.2), and d4~N (1.5, 0.2). The difficulty parameter was identified in a similar manner in the study by Kim and Lee (2004).

Data Analysis

Two test forms with and one test form without parameter drift were equated with the baseline test using the CINEG design. Test equating process was carried out using the SL (Stocking & Lord, 1982) method that accounted the difficulty and discrimination parameters

(Hambleton, Swaminathan & Rogers, 1991). Two test characteristic curves that were generated using X and Y forms were compared using this method (Kim, Harris & Kolen, 2010). The equation generated by the SL method is shown below:

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{i_j}\left(\theta_{j_i}; a_{j_i}, b_{j_i}, c_{j_i}\right) - p_{i_j}(\theta_{j_i}; \frac{a_{i_j}}{A}, Ab_{i_j} + B, c_{i_j} \right]^2 \qquad 1$$

Where A ise slope, B is constant, $p_{i_j}\left(\theta_{j_i}; a_{j_i}, b_{j_i}, c_{j_i}\right)$ is item characteristic function and $p_{i_j}(\theta_{j_i}; \frac{a_{i_j}}{A}, Ab_{i_j} + B, c_{i_j})$ is equated item characteristic function.

$$H_{crit} = \sum_{i} Hdiff(\theta_i) \qquad 2$$

This study used the R programming language (R Development Core Team, 2021) for data generation, calibration, and true score equation. Item and ability parameter estimations were carried out using package ‘irtplay’ (Lim, 2020), and true score equating was carried out using package ‘plink’ (Weeks, 2010). The root mean squared error (RMSE) value for equating errors was calculated using R, and it displayed the combination of bias (systematic error indicator) and conditional standard error (Kolen & Brennan, 2004).

The RMSE value was calculated using the equation below where θ stands for true ability, θ* stands for estimated true ability, and f stands for ability frequency (Deng & Monfils, 2017; Keller & Keller, 2011). Since this was a simulation study based on replications, the means of RMSE values were calculated.

$$RMSE = \sqrt{\frac{\sum_i f_i(\theta^* - \theta)^2}{\sum_i f_i}} \qquad 3$$

Factorial analysis of variance (ANOVA) was used to determine whether RMSE values displayed significant differences according to the manipulated conditions in the study for independent samples. In addition, effect size value was calculated.

Results

The equating process was carried out using the SL method under different simulation conditions that were identified for the study. RMSE values were calculated in the equating process using estimated and true abilities. The results are displayed in Figure 3. In addition, results belonging to each simulation condition are provided in Appendix 1.

Figure 3 displays that as a result of equating Form 4 (without parameter drift) to Form 1 (baseline test form), equating errors that approached 0.05 were obtained. An increase in sample size in tests without IPD resulted in an equating error that approached 0. Test length and changes in common item ratio yielded similar results in tests without IPD.

Errors generated by equating the two tests with IPD (Form 2 and Form 3) to the baseline test (Form 1) varied according to IPD ratio. When the IPD ratio dropped from 30 to 20%, equating errors decreased under all conditions. Equating errors peaked for Form 3, which was generated by running repeated parameter drift on Form 2.

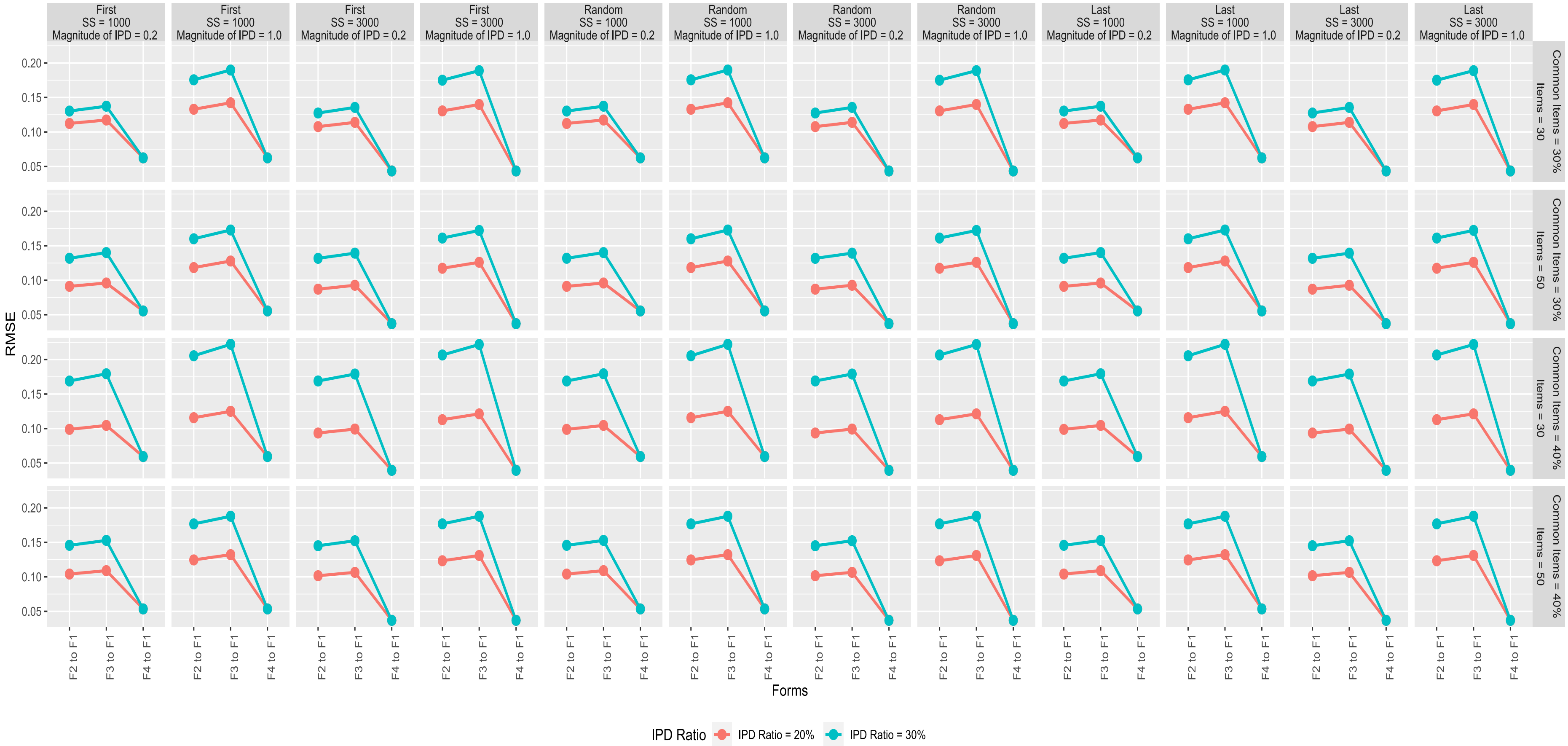

Figure 3. Equating errors (RMSE).

An overview of the graphs revealed an increase in IPD that resulted in an increase in equating errors, as expected. When sample size was increased from 1000 to 3000, errors either remained at the same level or slightly decreased. Common items located in the beginning, at the end, or distributed randomly yielded similar results in equating errors. An increase in common items when common items contained 30% IPD resulted in increased equating errors. When examined according to test length, an increase in test length caused equating errors to either remain at the same level or decrease. However, an increase in test length when common item ratio was 40% and IPD ratio was 20% caused equating errors to increase. Increasing test length from 30 to 50 when common item ratio was 40% and IPD ratio was 30% resulted in a significant decrease in equating errors.

Factorial analysis of variance (ANOVA) was used for independent samples to determine whether conditions of equating tests with and without IPD, common item ratio, ratio of items with IPD, IPD size, location of common items in tests, sample size, test length, and combination of these conditions displayed significant differences.

When the results of factorial ANOVA for independent samples were examined, errors generated by equating Forms 2, 3, and 4 to Form 1 displayed significant differences ($F_{(2,287956)}$=181067.21, p<0.05, η^2=0.56). The impact of test forms on test equating errors were significantly high. Partial eta squared value of 0.14 and above indicates a high level of impact (Cohen, 1977). As a result of dichotomous comparisons using the Bonferroni correction, errors generated by equating Form 4, without IPD, to Form 1 (p<0.05), Form 2, with IPD, to Form 1 (p<0.05), and Form 3, with IPD, to Form 1 (p<0.05) displayed significant differences. Mean of errors generated by equating Form 4 to Form 1 ($\bar{X}$=0.048) was significantly lower than that generated by equating Form 2 to Form 1 ($\bar{X}$=0.136), which was even lower than that generated by equating Form 3 to Form 1 ($\bar{X}$=0.145). Evidently, errors generated by equating Forms 2 and 3, with IPD, to the baseline test, Form 1, displayed significant differences (p<0.05). Test equating errors did not display statistically significant differences ($F_{(2,287956)}$=0.00, p>0.05, η^2=0.00) depending on the location of common items in tests.

According to Cohen (1977), partial eta squared value of 0.01 indicates low impact. Mean of equating errors obtained ($\bar{X}$=0.107) when common item ratio was 30% was significantly lower than the mean obtained ($\bar{X}$=0.113) when common item ratio was 40%. Mean of equating errors obtained ($\bar{X}$=0.114) when sample size was 1000 was significantly higher than the mean obtained ($\bar{X}$=0.106) when sample size was 3000. Similarly, mean of equating errors obtained ($\bar{X}$=0.115) when test length was 30 was significantly higher than the mean obtained ($\bar{X}$=0.105) when test length was 50. The impact of IPD size on test equating errors was at medium-level. Partial eta squared value of 0.06 indicates medium impact (Cohen, 1977). Mean of equating errors obtained ($\bar{X}$=0.100) when IPD size in multiple-choice items was 0.2 was significantly lower than the mean obtained ($\bar{X}$=0.120) when IPD size was 1.0. Aside from these, whether combinations of above-mentioned variables led to significant differences on test equating errors was also examined. The combinations with low and high effect size that displayed significant differences are discussed below.

Test forms and IPD size ($F_{(2,287956)}$=5236.27, p<0.05, η^2=0.04), test forms and IPD ratio ($F_{(2,287956)}$=14803.90, p<0.05, η^2=0.09), IPD ratio and common item ratio ($F_{(1,287956)}$=4341.53, p<0.05, η^2=0.02), test forms and sample size ($F_{(2,287956)}$=1488.46, p<0.05, η^2=0.01) as well as test forms and common item ratio ($F_{(2,287956)}$=958.47, p<0.05, η^2=0.01) display statistically significant differences in test equating errors. While the effects of test forms and IPD ratio on equating errors are above medium-level, the effect of test forms and IPD size on equating errors are just below medium-level. The effects of IPD ratio and common item ratio, test forms and sample as well as test forms and common item ratio on test equating errors are low.

Discussion

This study examined the impact of IPD on test equating when common items were in a mixed format. By manipulating the test length, sample size, common item ratio, location of common items in tests, IPD size, and items with IPD in common items, their impacts on equating errors were examined. It was found that IPD increased equating errors in test equating processes, and the errors displayed a statistically significant difference.

When the results were examined according to test length and sample size, equating errors decreased or were close to each other when the number of items in the test and sample size individually increased though the effects of these factors on test equating errors were significant yet low. An increase in test length when IPD ratio was 20% and common item ratio was 30% resulted in decreased equating errors. Lee and Geisinger (2019) obtained similar findings in their study that selected scales at different lengths that were rated polytomous in TIMMS. They emphasized that a decrease in the number of items in the scale resulted in an increased mean error, and the factor that most impacted RMSE values was scale length. An independent increase in test length and sample size when items with IPD were present, thus, caused equating errors to either remain the same or decrease. An increase in sample size when IPD was not present caused equating errors to approach zero.

An increase in the number of items with IPD in common items when common item ratio in the test increased resulted in a statistically significant increase in equating errors, however, the effect of this factor on equating errors was low. Han et al. (2012) reported that a decrease in the number of common items that were similarly scaled resulted in increased standard equating errors. With regards to IPD ratio in common items, a decrease in the number of items with IPD resulted in decreased equating errors. These differences were statistically significant, and their effects on test equating errors were high. In other words, the occurrence of IPD in common items resulted in increased equating errors. Having reached similar results, Li (2012) stated that the anchor test length and number of polytomous items with IPD were the two factors that affected equating results, and an increase in anchor test length resulted in a decrease in scaling and equating errors. Additionally, one polytomous item with IPD among common items reflected lower RMSE values compared to when the number of such items was two. Furthermore, Han and Wells (2007) stated that even when 10% of common items had IPD, results of test equating processes were significantly affected. Jimenez (2011), who compared the performances of SL and Haebara methods under various IPD conditions by using the CINEG design, stated that an increase in the number of items with IPD in common items, in particular, resulted in a significant increase in error values when non-equivalent groups were present. Huang and Shyu (2003) stated that when IPD occurred in half of the common items, particularly, pass/fail scores were significantly affected. In a similar study, Hu et al. (2008) stated that the exclusion of items with IPD resulted in fewer equating errors but led to other equating errors (Miller & Fitzpatrick, 2009) and increased RMSE values (Lee & Geisinger, 2019). The exclusion of polytomous items with IPD decreased the representative power of the test in general, and hence, polytomous items with IPD should be excluded from tests after identifying the extent of their impact on IRT equating results (Li, 2012). For both polytomous and dichotomous items, their invariance to IPD should be inspected, particularly in the successive administration of tests. While Kolen and Brennan (2004) recommended that common items be inspected for IPD so as not to introduce differences among groups to be equated, Babcock and Albano (2012) recommended that these inspections be carried out at regular intervals using IPD identification methods.

While a combined increase in the ratios of common item and IPD in common items resulted in significantly increased equating errors, the location of common items in tests yielded close results in terms of test equating errors and equating errors as this factor was found to be statistically insignificant. The location of common items when items with IPD were present did not have an effect on equating errors under the conditions of this study. However, in their study, Meyers et al. (2009) stated that a change in the location of items in tests significantly affected item difficulty in the Rasch model to the advantage of successful students and detriment of unsuccessful students. Similarly, in their study that used the SL equating method, Meyers et al. (2012) stated that a change in location of items in administration of tests resulted in more biases and errors, and item parameters could not be estimated consistently. This resulted in students being incorrectly categorized. However, these studies did not assess whether items with IPD were present, and only examined the impact of the change in location of these items in tests. Therefore, when changing the location of common items in tests, both the effects of IPD, if present, and the change of location condition on test equating should be examined. Additionally, since a high common item ratio resulted in increased equating errors, it is recommended to use common items ratios that are established by studies in literature.

When the results were examined according to IPD size, an increase in IPD size resulted in increased test equating errors, and this impact was at medium-level. Similarly, Li (2012), who examined the occurrence of IPD when polytomous items were present among common items, stated that an increased IPD size negatively affected scaling coefficients and equating accuracy, and that these items should be excluded from the test as long as content validity was not significantly affected (Li, 2012; Miller & Fitzpatrick, 2009). Rupp and Zumbo (2003a, 2003b) stated that IPD size at low levels did not affect ability estimations. In an equating study, Chen (2013) stated that an increase from 0.20 to 0.40 in parameter -b resulted in a decreased accuracy of theta estimations. Since an increased IPD size implied a drift in real parameter values of items, it is a factor that negatively affected the parameter invariance characteristic of equating, and, thus, increased the equating errors. Therefore, an inspection of parameter invariance is recommended before carrying out test equating processes.

To accurately estimate the performance and development levels of students in an accountable education system, parameters of common items in tests should maintain their invariance characteristic, particularly for test equating. Since the CINEG design uses common items for equating processes, the change in parameter values of these items in administration of tests may result in increased test equating errors, and, thus, lead to an incorrect estimation of individuals' abilities. While no exact rule is in place for dealing with items with IPD, common items should be inspected at regular intervals to determine whether they display IPD occurrences to minimize test equating errors. According to items' IPD conditions, a final decision should be made to either exclude the item from the test or revise the relevant items to maintain test validity. Regardless of the decision about items with IPD, their impact on the content validity of the test in general should also be assessed.

As in many simulation studies, this study was limited by specific conditions. Taking into account the results obtained from these conditions, some recommendations may be recommended for further studies. The impact of IPD on test equating could be examined on large-scale examinations that contain real data. The generalizability of this study is limited to previously examined test length, sample size, common item ratio, location of common items in tests, IPD size, and ratio of items with IPD in common items. Further studies may examine the impact of IPD on test equating by taking into account other conditions. Studies in literature mostly examine the impact of IPD on test equating in multiple-choice items. However, there are only few studies that examine this impact with open-ended items that are frequently applied to measure students' high-level cognitive abilities. This study used at most three polytomous items with IPD among the common items. Therefore, the impact of increase in the number of polytomous items with IPD on test equating results should be examined.

References

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

Arce-Ferrer, A. J., & Bulut, O. (2017). Investigating separate and concurrent approaches for item parameter drift in 3PL item response theory equating. *International Journal of Testing*, *17*(1), 1-22. doi:10.1080/15305058.2016.1227825

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*(7), 565–580. doi:10.1177/0146621612455090

Bulut, O., & Sunbul, O. (2017). Monte Carlo simulation studies in Item Response Theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 266-287. doi:10.21031/epod.305821

Bock, D. B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285. http://www.jstor.org/stable/1434961

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307-333). New York: Routledge/Taylor & Francis.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets* (Doctoral Dissertation, University of Maryland). Retrieved from https://drum.lib.umd.edu/handle/1903/8843

Carsey, T. M., & Harden, J. J. (2014). *Monte Carlo simulation and resampling methods for social sciences*. doi:10.4135/9781483319605

Chen, Q. (2013). *Remove or keep: Linking items showing item parameter drift* (Unpublished doctoral dissertation). Michigan State University, Michigan.

Chon, K. H., Lee, W.-C., & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests* (Research Report 26). Iowa: Center for Advanced Studies in Measurement and Assessment.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

DeMars, C. (2010). *Item response theory: Understanding statistics, measurement*. New York: Oxford University.

Demirus, K. B., & Uysal, İ. (2016, September). *The impact of item parameter drift on test equating*. Paper presented at the V. National Congress on Measurement and Evaluation in Education and Psychology, Antalya.

Deng, W., & Monfils, R. (2017). *Long-term impact of valid case criterion on capturing population-level growth under item response theory equating* (Research Report 17-17). Retrieved from https://doi.org/10.1002/ets2.12144

Felan, G. D. (2002, February). *Test equating: Mean, linear, equipercentile and item response theory*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin.

Gaertner, M. N., & Briggs, D. C. (2009). *Detecting and addressing item parameter drift in IRT test equating contexts*. Boulder, CO: Center for Assessment.

Guo, R., Zheng, Y., & Chang, H. (2015). A stepwise test characteristic curve method to detect item parameter drift. *Journal of Educational Measurement, 52*(3), 280-300. doi:10.1111/jedm.12077

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

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates* (Doctoral dissertation, University of Massachusetts Amherst). Retrieved from http://scholarworks.umass.edu/dissertations/AAI3325324

Han, K. T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (Research Report 11-02). Reston, Virginia: Graduate Management Admission Council.

Han, K., & Wells, C. S. (2007, April). *Impact of differential item functioning (DIF) on test equating and proficiency estimates*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Han, K. T., Wells, C. S., & Hambleton, R. K. (2015). Effect of adjusting pseudo-guessing parameter estimates on test scaling when item parameter drift is present. *Practical Assessment, Research, and Evaluation, 20*, 16. doi:10.7275/jyyy-wp74

Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education, 25*(2), 97-117. doi:10.1080/08957347.2012.660000

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. doi:10.1177/0146621602026001001

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311–333. doi:10.1177/0146621606292215

Huang, C. Y., & Shyu, C. Y. (2003, April). *The impact of item parameter drift on equating*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Jimenez, F. A. (2011). *Effects of outlier item parameters on IRT characteristic curve linking methods under the common-item nonequivalent groups design* (Unpublished master's thesis). University of Florida, Florida.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement, 71*(2), 362–379. doi:10.1177/0013164410375111

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419. doi:10.1177/0146621612446170

Kilmen, S. (2010). *Comparison of equating errors estimated from test equation methods based on item response theory according to the sample size and ability distribution* (Unpublished doctoral dissertation). Ankara University, Ankara.

Kim, S., Harris, D. J., & Kolen, M. J. (2010). Equating with polytomous item response models. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 257-291). New York: Routledge.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (Research Report 2004-5). Iowa: ACT.

Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53-76. doi:10.1111/j.1745-3984.2006.00004.x

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*(1), 36–53. http://www.jstor.org/stable/25651535

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scalling and linking*. New York: Springer.

Lee, H., & Geisinger, K.F. (2019). Item parameter drift in context questionnaires from international large-scale assessments. *International Journal of Testing, 19*(1), 23-51. doi:10.1080/15305058.2018.1481852

Li, X. (2008). An investigation of the item parameter drift in the examination for the certificate of proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 1-28.

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report 12-09). New Jersey: Educational Testing Service.

Lim, H. (2020). *irtplay: Unidimensional item response theory modeling* (Version 1.6.2) (Computer software). Retrieved from https://CRAN.R-project.org/package=irtplay

Marengo, D., Miceli, R., Rosato, R., & Settanni, M. (2018). Placing multiple tests on a common scale using a post-test anchor design: Effects of item position and order on the stability of parameter estimates. *Frontiers in Applied Mathematics and Statistics, 4*, 50. doi:10.3389/fams.2018.00050

McCoy, K. M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment* (Unpublished doctoral dissertation). Champaign, IL: University of Illinois.

Meng, Y. (2012). *Comparison of Kernel equating and item response theory equating methods* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3518262)

Meng, H., Steinkamp, S., & Matthews-Lopez, J. (2010, May). *An investigation of item parameter drift in computer adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38–60. doi:10.1080/08957340802558342

Meyers, J. L., Murphy, S., Goodman, J., & Turhan, A. (2012, April). *The impact of item position change on item parameters and common equating results under the 3PL model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, B.C.

Michaelides, M. P. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology, 1*, 167. doi:10.3389/fpsyg.2010.00167

Miller, G. E., & Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educational and Psychological Measurement, 69*(3), 357-368.

Mooney, C. Z. (1997). *Monte Carlo simulation* (Series No. 07-116). doi:10.4135/9781412985116

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. doi:10.1177%2F014662169201600206

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61-78. doi:10.1080/08957340802558367

R Development Core Team. (2021). *R: A language and environment for statistical computing* (Cersion 4.0.5) (Computer software). R Foundation for Statistical Computing.

Rupp, A. A., & Zumbo, B. D. (2003a, April). *Bias coefficients for lack of invariance in unidimensional IRT models*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*(3), 264-276. doi: 10.11575/ajer.v49i3.54984

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*(1), 65-88. doi:10.1080/08957340701796415

Skaggs, G., & Lissitz, R. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529. doi:10.2307/1170343

Stahl, J. A., & Muckle, T. (2007, April). *Investigating displacement in the Winsteps Rasch calibration application*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Stocking, M. L., & Lord, F. M. (1982). *Developing a common metric in item response theory* (Research Report 82-25-ONR). Retrieved from https://doi.org/10.1002/j.2333-8504.1982.tb01311.x

Sukin, T. M. (2010). *Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees* (Doctoral dissertation, University of Massachusetts Amherst). Retrieved from https://scholarworks.umass.edu/open_access_dissertations/301

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*(4), 329-346. doi:10.1111/j.1745-3984.2000.tb01090.x

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT* (Doctoral dissertation, Boston College). Retrieved from http://hdl.handle.net/2345/2370

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor item response theory approach. *Frontiers in Psychology, 7*, 270. doi:10.3389/fpsyg.2016.00270

Weeks, J. P. (2010). *plink: An R package for linking mixed-format tests using IRT-based methods* (Version 1.5-1) (Computer software). *Journal of Statistical Software, 35*(12), 1-33. doi:10.18637/jss.v035.i12

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77–87. doi:10.1177/0146621602261005

Wollack, J. A., Sung, H. J., & Kang, T. (2005, April). *Longitudinal effects of item parameter drift*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Appendix 1. RMSE Values

Ratio of IPD	Percentage of Common Items	Link	Location of IPD Items																							
			First								Random								Last							
			Sample Size																							
			1000				3000				1000				3000				1000				3000			
			The Number of Items																							
			30		50		30		50		30		50		30		50		30		50		30		50	
			Magnitude of IPD																							
			0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0	0.2	1.0
%20	%30	Form2 to Form1	0.112	0.133	0.091	0.118	0.108	0.130	0.087	0.117	0.112	0.133	0.091	0.118	0.108	0.130	0.087	0.117	0.112	0.133	0.091	0.118	0.108	0.13	0.087	0.117
		Form3 to Form1	0.117	0.142	0.096	0.128	0.114	0.140	0.093	0.126	0.117	0.142	0.096	0.128	0.114	0.140	0.093	0.126	0.117	0.142	0.096	0.128	0.114	0.14	0.093	0.126
		Form4 to Form1	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037
	%40	Form2 to Form1	0.099	0.116	0.104	0.125	0.094	0.113	0.102	0.123	0.099	0.116	0.104	0.125	0.094	0.113	0.102	0.123	0.099	0.116	0.104	0.125	0.094	0.113	0.102	0.123
		Form3 to Form1	0.105	0.125	0.109	0.132	0.099	0.121	0.106	0.131	0.105	0.125	0.109	0.132	0.099	0.121	0.106	0.131	0.105	0.125	0.109	0.132	0.099	0.121	0.106	0.131
		Form4 to Form1	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037
%30	%30	Form2 to Form1	0.13	0.176	0.132	0.16	0.127	0.175	0.132	0.161	0.13	0.176	0.132	0.16	0.127	0.175	0.132	0.161	0.13	0.176	0.132	0.16	0.127	0.175	0.132	0.161
		Form3 to Form1	0.137	0.19	0.14	0.173	0.136	0.189	0.139	0.172	0.137	0.19	0.14	0.173	0.136	0.189	0.139	0.172	0.137	0.19	0.14	0.173	0.136	0.189	0.139	0.172
		Form4 to Form1	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037	0.062	0.062	0.055	0.055	0.043	0.043	0.037	0.037
	%40	Form2 to Form1	0.169	0.205	0.146	0.177	0.169	0.207	0.145	0.177	0.169	0.205	0.146	0.177	0.169	0.207	0.145	0.177	0.169	0.205	0.146	0.177	0.169	0.207	0.145	0.177
		Form3 to Form1	0.179	0.222	0.153	0.188	0.179	0.222	0.152	0.188	0.179	0.222	0.153	0.188	0.179	0.222	0.152	0.188	0.179	0.222	0.153	0.188	0.179	0.222	0.152	0.188
		Form4 to Form1	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037	0.059	0.059	0.053	0.053	0.039	0.039	0.037	0.037