Psychometric Consequences of Subpopulation Item Parameter Drift

Educational and Psychological
Measurement

2016, Vol. 77(1) 143–164

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416643369

journals.sagepub.com/home/epm

Anne Corinne Huggins-Manley¹

Abstract

This study defines subpopulation item parameter drift (SIPD) as a change in item parameters over time that is dependent on subpopulations of examinees, and hypothesizes that the presence of SIPD in anchor items is associated with bias and/or lack of invariance in three psychometric outcomes. Results show that SIPD in anchor items is associated with a lack of invariance in dimensionality structure of an anchor test, a lack of invariance in scaling coefficients across subpopulations, and a lack of invariance in ability estimates. It is demonstrated that these effects go beyond what can be understood from item parameter drift or differential item functioning.

Keywords

item parameter drift, invariance, differential item functioning, fairness

Introduction

As a matter of test fairness and scale stability, item parameters are expected to be invariant across different subpopulations of examinees and across different administrations of test items. A lack of invariance in item parameters across subpopulations of examinees is defined as differential item functioning (DIF; Holland & Thayer, 1988). Investigations of DIF have been commonplace in operational testing programs for several decades (Penfield & Camilli, 2007), and research on DIF in relation to various testing concerns continues to permeate the field of psychometrics to the present year (e.g., Andrich & Hagquist, 2015; Li & Wang, 2015). A lack of invariance in item parameters across test administrations is defined as item parameter drift (IPD),

¹University of Florida, Gainesville, FL, USA

Corresponding Author:

Anne Corinne Huggins-Manley, College of Education, University of Florida, 1215 Norman Hall, Gainesville, FL 32611, USA.

Email: ahuggins@coe.ufl.edu

and can be conceptualized as a special case of DIF in which the subpopulations are defined by administration occasion (Goldstein, 1983). IPD indicates a lack of scale stability due to item parameter changes over time. IPD research has focused on how IPD impacts various properties of test scores and scales (e.g., Han, Wells, Sireci, 2012; Wells, Subkoviak, & Serlin, 2002), methods for detecting IPD (e.g., DeMars, 2012; Wells, Hambleton, Kirkpatrick, & Meng, 2014), and applications of IPD detection in particular testing contexts (e.g., Krause, 2012; Giordano, Subhiyah, & Hess, 2005).

Figure 1 displays various b parameters (i.e., item difficulty parameters) from a single test item exposed to two subpopulations of examinees over two test administrations. For simplicity, b parameters from an item response theory (IRT) model are used in Figure 1, although other item parameters could be used in place of b parameters. DIF methods compare item parameters across subpopulations within a single test administration, as shown in the top table of Figure 1. IPD methods compare item parameters across administrations ignoring subpopulations, as shown in the middle table of Figure 1.

This study hypothesizes that subpopulation item parameter drift (SIPD) may capture some specifics of item parameter invariance or lack thereof that are overlooked by IPD and DIF. SIPD compares item parameters across administrations within a single subpopulation, as shown in the bottom table of Figure 1. For example, evidence of SIPD would be present if an item's b parameter for English language learners was found to drift from one test administration to the next while the b parameter for non-English language learners did not drift. If all subpopulations' b parameters drifted in the same direction and magnitude, then SIPD would be equivalent to IPD. However, if SIPD differed across subpopulations (i.e., differential SIPD) then its presence would be conceptually distinguished from both IPD and DIF. One can view this differential SIPD as an interaction between IPD and subpopulations because differential SIPD is present if the change in b parameters over time (IPD) is dependent on subpopulation. Alternatively, one could view differential SIPD as an interaction between DIF and test form because differential SIPD is present if the difference in b parameters across subpopulations is dependent on test form.

Formally, differential SIPD in b parameters is defined as

$$(b_{F,j} - b_{F',j}) \neq (b_{F,j'} - b_{F',j'}), \quad (1)$$

where F is a test form, F' is a different test form, j is a subpopulation, and j' is a different subpopulation. Equation (1) is stated as, "the difference in b parameters across forms for one subpopulation is unequal to the difference in b parameters across forms for a different subpopulation." A similar definition could be applied to other item parameters of interest.

Given certain subpopulation sample sizes and varying magnitudes of IPD and DIF across subpopulations, it is possible for SIPD to exist independent of the existence of meaningful DIF and IPD. In addition, because SIPD is a phenomenon unique from IPD and DIF, it may be associated with a distinct set of psychometric issues that are

Figure 1. Conceptualization of a single item's b parameters that are compared under DIF, IPD, and SIPD definitions. DIF = differential item functioning; IPD = item parameter drift; SIPD = subpopulation item parameter drift.

unexplained by IPD or DIF alone. This study examines the effect of differential SIPD on multiple psychometric outcomes, and explores the notion that its effects are distinguishable from those explained by DIF and IPD.

Background and Research Questions

The hypothesis that differential SIPD may be important to define and examine originated in a recent study on DIF in anchor items used to equate test forms. Huggins (2014) examined the effects of anchor item DIF on population invariance of IRT true score equating and found that population invariance of equating was compromised to various degrees when DIF varied across forms in a differential manner across subpopulations. Consistent DIF across forms was unable to explain the effect on equating invariance. The study did not set out to examine differential SIPD as defined above or even IPD in general. Yet the results uncovered the reality that if item parameters

drift differently over time for one subpopulation of examinees as compared to another, equating dependence is an expected outcome. This poses concerns for score equity, as defined by equated, reported scores that differ when subpopulations are treated separately in measurement analysis procedures as compared to when they are treated together (Dorans, 2004). Because SIPD was not the focus of the study, there were limited types of SIPD that were explored and a focus on only one type of psychometric outcome. It was concluded that more needs to be investigated on this topic.

But a continuation of the phenomena of SIPD is only warranted if indeed subpopulation-level item parameter estimates can be expected to drift differently in operational testing programs. Concerning IPD, there is evidence that changes in curricular emphasis over time can cause item parameters to drift. Goldstein (1983) and Mislevy (1982) suggested that curriculum changes may be a cause of observed IPD. Sykes and Fitzpatrick (1992) hypothesized that changes in curriculum emphasis were a key explanatory factor for IPD patterns in an operational testing program. Bock, Muraki, and Pfeifferberger (1988) found that the relative direction of drift in a large scale physics test was associated with changed emphasis in curriculum over time. Deng and Melican (2009) supposed that the relationship they uncovered between content and drift may have been due to curriculum emphasis changes.

From this, one can hypothesize that if two subpopulations are exposed to different types of curricula and/or have different emphases within those curricula, then differential SIPD may result when those curricula or emphases change between test administrations. For example, curricula and the pace of curricula used for students with disabilities show many changes over time that are not necessarily aligned with changes or lack of changes in curricula for students without disabilities (e.g., Karger & Hitchcock, 2003). English language learners may be exposed to modified curricula that change at different times and in different ways than the curricula of non-English language learners (e.g., Rossell, 2003). In these circumstances, the effect of curricula emphasis on IPD would be inconsistent across subpopulations. These examples are specific to K-12 curricula and statewide tests of accountability, but similar examples could be made for other large-scale tests. For example, it can be assumed that law students from across the nation taking the high-stakes Multistate Bar Exam would be exposed to different curricula in their respective university programs. It is unreasonable to expect that changes in emphasis of law degree curricula across different universities and states operate in tandem. Hence, some differential SIPD for subpopulations based on university or state would be expected. Similarly, with international tests such as the Trends in International Mathematics and Science Study (TIMSS), one would expect that students from different countries are exposed to different curricula that are changing in different ways over time.

Another reason that differential SIPD may occur in operational tests is related to opportunity to learn factors. Just as IPD can result from opportunity to learn changes across test administrations (Sukin, 2010), differential SIPD could result from opportunity to learn changes that are specific to subpopulations. Even when the intended

curricula are the same across various subpopulations of students, different opportunities to learn that curricula can impact between-group measurement on summative large-scale tests (e.g., Snyder, Mereku, Amedahe, Etsey, & Adu, 2013). This may be particularly problematic in the current U.S. K-12 landscape in which many states are defining low-achieving students as a subpopulation of interest in high-stakes measurement (Ujifusa, 2014). These students often display high mobility rates (e.g., Alexander, Entwisle, & Dauber, 1996), are associated with high absentee rates (e.g., Baxter, Royer, Hardin, Guinn, & Devlin, 2011), and have teachers with less experience and/or specialized training (e.g., Rennie Center of Education Research and Policy, 2013), all of which can affect exposure to classroom material and learning opportunities. Again, the differential experiences in the classroom relate to different curricula exposure across time and hence support the expectation that differential SIPD across one or more subpopulations of test examinees may exist on various tests.

In the instance that SIPD does exist, there are several reasons that it is important to locate it and understand the effects it can have on various test outcomes. First, it may be that practical problems extend from SIPD similar to those that have been found within IPD studies, such as the tendency to affect examinee ability estimates (Wells et al., 2002), relative performance of various linking methods (Chen, 2013; Han et al., 2012), and performance of vertical scaling methods (Ye & Xin, 2014). All of these issues would be problematic for ensuring that subpopulations are measured fairly and consistently across groups. Second, one of the hallmark features of *No Child Left Behind* (NCLB; 2001) was the requirement to disaggregate test scores by subpopulations of students, and this feature of score reporting has been retained in all revisions to NCLB (U.S. Senate, 2015) as well as the most current Congress- and President-approved draft of the *Every Student Succeeds Act* (Education Week, 2015; Klein, 2015). The effect of SIPD needs to be understood to determine its impact on our ability to disaggregate scores under these testing policies while assuming that the scores are comparable across subpopulations. Indeed, national agencies mandate test developers to ensure comparability of reported subpopulation test scores (Joint Committee on Testing Practices, 2004; U.S. Department of Education, 2009), and SIPD may have an impact on this comparability. Third, even when subpopulation level scores are not reported and/or used for high-stakes decisions, it is a matter of test fairness that invariant relationships exist in the psychometric properties of operational tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), and the presence of differential SIPD would violate these principles if it were found to relate to a lack of invariance in ability estimates and other psychometric outcomes.

Finally, research on population invariance of equating has often focused on the quantification of population invariance and the detection of it on operational tests (e.g., Dorans & Liu, 2009; Yi, Harris, & Gao, 2008), and an outcome of this body of research is the established need to understand more about the relationships between the same construct property of equating, multidimensionality and equating invariance (Yang & Gao, 2008). Huggins (2014) began addressing this need by connecting DIF

as a manifestation of differential construct assessment to population invariance of equating. However, dimensionality and other psychometric mechanisms that connect DIF to equating invariance have yet to be explored and the connection of DIF to equating invariance in that particular study was found to be dependent on how the DIF was manifested over time. Yet, the intricacies of that finding were not further elaborated. This study defines SIPD as a special form of DIF and IPD that is expected not only to explain equating invariance, but to connect that explanation to issues of anchor test dimensionality, scaling coefficients, and ability estimates that arise from SIPD.

The purpose of this study is to examine various psychometric consequences of SIPD and to determine if those consequences are differentiated from those that can be explained by IPD and DIF. This study introduces various manifestations of SIPD into anchor items drawn from a simulated environment, calibrates and scales scores from two test forms both across and within subpopulations, and examines the impact of the SIPD manifestations on bias and lack of invariance (LOI) with respect to multiple psychometric outcomes. The research questions are (a) Under what conditions does SIPD impact bias and LOI of the anchor test dimensionality? (b) Under what conditions does SIPD impact bias and LOI of scaling coefficients? (c) Under what conditions does SIPD impact bias and LOI of ability estimates? (d) Do the answers to research questions *a*, *b*, and *c* signify that SIPD has consequences above and beyond those that can be explained by DIF and IPD?

Method

A simulation was conducted in the R environment (R Development Core Team, 2011), with item parameters and test forms based on a Grade 4 mathematics Florida Comprehensive Assessment Test (Human Resources Research Organization & Harcourt Assessment, 2006). The two simulated test forms (i.e., Form 1 and Form 2) each consisted of 48 dichotomously scored items, 8 of which served as internal anchor items. True item parameters were based on estimated three-parameter logistic model (3PL; Birnbaum, 1968) item parameters from the operational test shown in Table 1. For the 40 nonanchor items, one fourth of each of the *a*, *b*, and *c* parameters were drawn from a uniform distribution bounded either by the minimum and the first quartile, the first quartile and the median, the median and the third quartile, or the third quartile and the maximum. The sampled *a*, *b*, and *c* parameters were then independently randomized across the 40 nonanchor items. The 8-item anchor test was simulated in an identical fashion.

Simulation Factors and Data Generation

Four factors were varied in the simulation. First, a differential SIPD factor was introduced by manipulating *b* parameters across two subpopulations of examinees in three ways: (a) the focal subpopulation displayed a positive, moderate *b* parameter drift

Table 1. Estimated Item Parameter Quartiles from a Grade 4 Florida Comprehensive Assessment Test.

Item parameter	Minimum	First quartile	Median	Third quartile	Maximum
a^a	0.36	0.70	0.88	1.02	1.43
b	-2.63	-1.14	-0.24	0.35	0.92
c	0.06	0.12	0.18	0.23	0.41

^aEstimated with a $D = 1.7$ scaling constant, which was then mirrored in the simulation.

from Form 1 to Form 2 while the other displayed no drift (i.e., $b_{foc,F1} < b_{foc,F2}$; $b_{ref,F1} = b_{ref,F2}$), (b) the focal subpopulation displayed a positive, moderate b parameter drift while the reference subpopulation displayed a negative, moderate b parameter drift (i.e., $b_{foc,F1} < b_{foc,F2}$; $b_{ref,F1} > b_{ref,F2}$), (c) both subpopulations displayed a positive b parameter drift but of different magnitudes, with the focal drift being large and the reference drift being moderate (i.e., $b_{foc,F1} < b_{foc,F2}$ [large]; $b_{ref,F1} < b_{ref,F2}$ [moderate]), and (d) no subpopulations displayed drift ($b_{foc,F1} = b_{foc,F2}$, $b_{ref,F1} = b_{ref,F2}$), serving as null conditions. The magnitudes for moderate and large drift were based on Educational Testing Service’s DIF magnitude definitions (Dorans & Holland, 1993), and the moderate amount was similar to those used in a recent IPD study (Han et al., 2012). The differences in b parameters were manipulated such that they were not fully confounded with within-form DIF, in contrast to the above referenced Huggins (2014) study. Note that all nonnull SIPD conditions satisfy Equation (1).

Second, the number of items with SIPD was varied as either one anchor item or three anchor items. Third, the relative sample sizes of the subpopulations were varied based on reported sample sizes from the operational test in Table 1: (a) equal sample sizes as observed in gender subpopulations (i.e., $n_{foc} = 5,000$; $n_{ref} = 5,000$), (b) slightly unequal sample sizes as observed between some racial subpopulations (i.e., $n_{foc} = 3,000$; $n_{ref} = 7,000$), and (c) extremely unequal sample sizes as observed between ELLs and non-ELLs, students with disabilities and students without disabilities, or standard curriculum students as compared to non-standard curriculum students (i.e., $n_{ref} = 1,000$; $n_{foc} = 9,000$). Last, a factor of relative mean ability between the subpopulations was varied on three levels: (a) no mean differences with theta distributed as $\theta \sim N(0, 1)$ for both subpopulations, (b) small mean differences with the focal subpopulation drawn from a distribution of $\theta \sim N(-0.25, 1)$, and large mean differences with the focal subpopulation drawn from a distribution of $\theta \sim N(-1, 1)$. These sample size and mean difference conditions were based on an amalgam of various subpopulation mean differences observed on the operational test in Table 1. The crossing of the four factors resulted in 63 unique conditions within the study and 100 iterations were completed within each condition.

Data Analysis and Simulation Outcomes

Within-form data were calibrated both within and across subpopulations with the 3PL model by batching BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) through R. The Stocking and Lord (1983) method was used to scale ability estimates from Form 2 onto Form 1, with the use of the *equateIRT* package in R (Battaui, 2014). This scaling method was chosen based on the findings of Huggins (2014) in which it performed better than other scaling methods in the presence of DIF in anchor items.

IPD and DIF have been associated with both bias and lack of invariance (LOI) in various studies (e.g., Han et al., 2012; Rupp & Zumbo, 2006). Hence, both bias and LOI were examined in this study with respect to three different psychometric outcomes: dimensionality of the anchor test, scaling coefficients, and ability estimates.

Dimensionality of the Anchor Test. A unidimensional confirmatory factor analysis (CFA) was fit to the item responses on the anchor test (i.e., response vectors on the eight anchor items with data collapsed across both forms) in Mplus (Muthén & Muthén, 2012) using weighted least squares estimation with adjusted means and variances. First, the CFA was implemented on the population to assess if the anchor test had the same dimensionality across forms when SIPD was present as compared to when SIPD was not present, with differences between these two types of conditions considered evidence of bias. Next, the CFA was implemented separately for the focal and reference subpopulations to explore invariance in the dimensionality structure. For all CFA outcomes, fit was evaluated with the root mean square error of approximation (RMSEA), the comparative fit index (CFI) and the Tucker Lewis index (TLI). Graphical procedures were used to condition the outcomes on the study factors. No modeling procedures were used on the fit index outcomes due to extreme non-normality of fit index distributions within conditions and the clarity of the visual results.

Scaling Coefficients. Bias in the scaling constants was estimated as

$$RMSE_{\hat{B}} = \sqrt{\frac{\sum_{iterations} (\hat{B}_{SIPD} - \hat{B}_{NULL})^2}{100}}, \quad (2)$$

where $\hat{B}_{SIPD}$ represents a B scaling coefficient estimated in a condition that had some form of SIPD, and $\hat{B}_{NULL}$ represents a B scaling coefficient estimated in an analogous simulation condition with no SIPD. Each condition had a single $RMSE$ for the B scaling coefficients, and an identical formula was used to estimate a single $RMSE$ for the A scaling coefficient in each condition. An analysis of variance (ANOVA) procedure was used to understand the main and two-way interaction effects of the study factors on scaling coefficient bias. LOI in the scaling coefficients was estimated as

$$RMSD_{\hat{B}} = \sqrt{\sum_j w_j (\hat{B}_j - \hat{B}_P)^2}, \quad (3)$$

where j is a subpopulation and hence $\hat{B}_j$ is a B scaling constant derived from calibrations and scalings conducted within a subpopulation, $\hat{B}_P$ is a B scaling constant derived from calibrations and scalings conducted across the population of examinees, and w_j is a weight based on the proportional sample size of the subpopulation to the population. Equation (3) was also applied to the A scaling constants. ANOVA models, correlations, and graphical procedures were applied to understand the effects of SIPD, IPD, and DIF on scaling coefficient LOI.

Ability Estimates. Ability estimates on Form 2 (i.e., $\hat{\theta}_{F2}$) were scaled to Form 1 both within and across subpopulations, such that

$$\hat{\theta}_{eF1} = A\hat{\theta}_{F2} + B, \quad (4)$$

where $\hat{\theta}_{eF1}$ is an estimated ability score from Form 2 that has been scaled to Form 1. Because unique simulees were drawn across all the conditions, bias in ability estimates was examined at the composite distribution level rather than at the individual level. The mean of the scaled ability distributions within each condition with SIPD were compared to the mean of the scaled ability distributions within the respective null condition with no SIPD. A mean difference in estimated ability was calculated as

$$\bar{\hat{\theta}}_{bias} = \bar{\hat{\theta}}_{SIPD} - \bar{\hat{\theta}}_{NULL}. \quad (5)$$

Because the estimated abilities in Equation (4) were on the same scale and originated from study conditions that were identical except for the presence or absence of SIPD, all mean differences can be attributed to the effects of the drift manipulated in the study. Therefore, $\bar{\hat{\theta}}_{bias}$ is interpreted as the bias in mean ability estimates. An ANOVA was used to understand the main and two-way interaction effects of each study factor on ability estimate bias. LOI of ability estimates was estimated as

$$RMSD_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^N (\hat{\theta}_{eF1,j} - \hat{\theta}_{eF1,P})^2}{N}}, \quad (6)$$

where $\hat{\theta}_{eF1,j}$ is a Form 2 ability score that was calibrated and scaled to Form 1 within subpopulation j , and $\hat{\theta}_{eF1,P}$ is a Form 2 ability score that was calibrated and scaled to Form 1 within the population. ANOVA models and graphical procedures were used to understand the effects of SIPD, IPD, and DIF on ability estimate LOI.

Results

Dimensionality Results

Dimensionality Bias. Figure 2 displays the RMSEA, CFI, and TLI results, where each data point in a graph is the mean outcome across iterations. The simulation conditions of mean differences between subpopulations and sample size were left out of the figure as their effects on the outcomes did not provide consistent, meaningful interpretations. The results show that conditions with no SIPD had the most acceptable fit, with RMSEA values close to 0 and CFI and TLI close to 1.00. When one subpopulation drifted, fit worsened. When both subpopulations drifted but in opposite directions, fit worsened more and particularly when three items had SIPD. The worst range of fit indices came about when both subpopulations drifted in the same direction but with different magnitudes, with more items of SIPD usually but not always resulting in worse fit. Note that this type of similar SIPD across subpopulations could be considered IPD. Also note that RMSEA values were always in an acceptable range (with RMSEA < .06 as a benchmark), and CFI and TLI values were below or very close to an acceptable range (with CFI/TLI > .95 as a benchmark). So when looking at dimensionality across subpopulations and comparing various SIPD conditions to null conditions, the results were largely within an interpretation of the anchor test displaying acceptable unidimensionality and the presence of worsening, if not poor, fit is better explained by IPD than SIPD. Indeed, further graphical inspections (not shown) demonstrated that the amount of IPD was the best predictor of dimensionality bias. As for the relationship between DIF and the fit indices, the analyses showed no clear relationship between true Form 1 or Form 2 DIF and unidimensional CFA model fit to the anchor test.

Dimensionality Lack of Invariance. Figure 3 shows the RMSEA, CFI, and TLI results for the reference subpopulation in the left column and for the focal subpopulation in the right column. Type of SIPD and number of items with SIPD are displayed, but the mean difference and sample size conditions are not included as they had inconsistent effects on the outcomes. The results show (a) subpopulation-level dimensionality of the anchor test is dependent on SIPD specific to the subpopulation being modeled, regardless of how the other subpopulation's item parameters drifted—for example, the RMSEA of the anchor test CFA for the reference subpopulation was equal across conditions with no SIPD and conditions in which only the focal subpopulation displayed SIPD; (b) more SIPD for the subpopulation being modeled resulted in a wider range of unidimensional model fit indices, with more SIPD defined by both the number of items with SIPD and the magnitude of that SIPD; and (c) the focal subpopulation, displayed worse fit than the reference subpopulation, and the fit indices when the focal SIPD were large were sometimes in the unacceptable range with respect to CFI and TLI. This relative subpopulation comparison shows LOI concerns with dimensionality based on SIPD type, size, and number of items. IPD had an inconsistent relationship to the subpopulation-level fit indices. While the best fit indices were associated with true IPD of zero, the less ideal fit indices were scattered across a

Figure 2. Type of SIPD and number of items with SIPD in relation to RMSEA, CFI, and TLI from the overall sample anchor test dimensionality assessment. SIPD = subpopulation item parameter drift; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

wide range of small to large true IPD. Similarly, all results indicated that DIF had no explanatory power for the patterns of anchor test dimensionality LOI. Specifically, many instances of low within-form DIF were associated with worse model fit indices as compared to other instances of higher within-form DIF, and vice versa.

Scaling Coefficient Results

An examination of bias in scaling coefficients was completed, followed by an examination of LOI in scaling coefficients. Neither bias nor LOI in the *A* scaling coefficients was present, so the results presented are only for the *B* scaling coefficients.

Figure 3. Type of SIPD and number of items with SIPD in relation to RMSEA, CFI, and TLI from the within subpopulation anchor test dimensionality assessment. SIPD = subpopulation item parameter drift; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

Figure 4. Graphical displays of the meaningful interactions from the various analyses of variance (ANOVAs).

Scaling Coefficient Bias. An ANOVA was completed for the outcome of B bias (i.e., $RMSE_B$ in Equation 2), and included the main effects and two-way interactions between all study factors. The interaction of number of items and type of SIPD had a very large effect. The top left graph in Figure 4 displays the interaction between number of items with SIPD and type of SIPD. It shows that the largest and only meaningful mean bias in B scaling coefficients was attributed to conditions in which three items drifted in the same direction across the subpopulations. This is an indication that IPD, rather than SIPD, is the explanatory factor for bias in the B scaling coefficients.

Scaling Coefficient Lack of Invariance. A similar ANOVA to the one exploring scaling coefficient bias was completed with $RMSD_B$ (see Equation 3) as the outcome. The interaction between number of items and type of SIPD were again the key

Figure 5. The strong relationship between differential SIPD and LOI in the B scaling coefficient. SIPD = subpopulation item parameter drift; LOI = lack of invariance.

explanatory factors of the outcome. However, the interaction shown in the top right corner of Figure 4 displays a core difference between these LOI results and the bias results. The conditions that explain LOI in the B scaling coefficient are those in which the subpopulations differ in item parameter drift (i.e., show differential SIPD). To further understand the findings, *observed* IPD, within-form DIF, and SIPD for each subpopulation were calculated from the estimated item parameters in each iteration of the study. Each was calculated as a simple difference in the relevant estimated b parameters for that index (see Figure 1), and summed across the three anchor items that were manipulated in the simulation. Weights were applied to groups where needed in these calculations. The observed values were aggregated across iterations within each condition. Because the ANOVA results showed that *differential* SIPD was important, the observed reference SIPD was subtracted from the observed focal SIPD to isolate focal SIPD from SIPD across the reference subpopulation. Figure 5 contains this differential SIPD in the x -axis, which can be read as “the degree to which the focal group’s $\hat{b}$ parameters drifted differently from the reference group’s $\hat{b}$ parameters.” This figure shows that estimated differential SIPD, along with true mean differences between groups, had very strong explanatory power for the observed LOI in the B scaling coefficient. Conversely, Figure 6 shows that IPD and

Figure 6. The lack of relationship between IPD/DIF and LOI in the B scaling coefficient. IPD = item parameter drift; DIF = differential item functioning; LOI = lack of invariance.

within-form DIF do not well explain LOI in the B scaling coefficient. Form 1 DIF is not shown as it had comparable results to Form 2 DIF.

Ability Estimate Results

Ability Estimate Bias. An ANOVA was completed for the outcome of $\hat{\theta}_{bias}$ (see Equation 5) and included the main effects and two-way interactions between all study factors. The interaction of number of items and type of SIPD had a very large effect and after exploring that interaction all other effects did not provide additional interpretative value. The graphic in the bottom left corner of Figure 4 displays the interaction. The largest bias in ability estimates was attributed to conditions in which three items drifted in the same direction across the subpopulations. This is an indication that IPD was the main explanatory factor for bias in ability estimates.

Ability Estimate Lack of Invariance. The ANOVA for $RMSD_{\hat{\theta}}$ (see Equation 6) showed once again that the interaction of number of items and type of SIPD was the driving explanatory effect of the outcome. This interaction shown in the bottom right corner of Figure 4 indicated that LOI in ability estimates was largest in study conditions in which multiple items displayed SIPD that differed in direction across the subpopulations, and it aligned with a 0.20 standard deviation difference between ability estimates based on subpopulation analyses as compared to population analyses. A follow up using observed focal SIPD minus observed reference SIPD (see explanation of this calculation in an above section) showed that differential SIPD along with subpopulation sample size differences had strong explanatory power with respect to the degree of LOI in ability estimates (see Figure 7). Very similar to the scaling coefficient LOI

Figure 7. The strong relationship between differential SIPD and LOI in ability estimates. SIPD = subpopulation item parameter drift; LOI = lack of invariance.

results, neither IPD nor within-form DIF could explain the variation in LOI in ability estimates.

When further explored, it was determined that the effect of sample size shown in Figure 7 was mostly due to the way $RMSD_{\hat{\theta}}$ was calculated. When $RMSD_{\hat{\theta}}$ was calculated separately for the focal and reference subpopulations it was clear that the focal subpopulation showed larger differences in proficiency estimates (see Equation 6). So when the focal subpopulation constitutes a larger proportion of the overall sample size, the $RMSD_{\hat{\theta}}$ results are larger.

Discussion and Implications

The research questions for this study centered around the impact of SIPD conditions on anchor test dimensionality, scaling coefficients, and ability estimates, while determining if those influences were distinguishable from that of IPD and DIF. The main finding was that IPD is the best explanatory factor of bias in the outcomes whereas differential SIPD across subpopulations is the best explanatory factor of LOI in the outcomes. Within-form DIF was unable to explain the outcomes because the effect of

one form's DIF on these outcomes was dependent on whether or not the other form had a different form of DIF, which defines SIPD.

When collapsing subpopulations and comparing conditions with some form of true SIPD to analogous conditions with no true SIPD, IPD explained much of the effects on bias in psychometric outcomes. With respect to dimensionality, IPD was helpful, although far from perfect, in explaining the differences in fit indices when data was collapsed across subpopulations. With respect to scaling coefficients and ability estimates, IPD in multiple items was by far the best predictor of bias. The results indicate that it is the aggregate drift in subpopulation item parameters that affects bias, and indeed SIPD can cancel out across subpopulations and result in no bias. This is similar to multidirectional IPD canceling out across items (Han et al., 2012), except in this case it is multidirectional SIPD canceling out across subpopulations. The result is that the total IPD across subpopulations explains bias in the various psychometric outcomes.

However, when we treat each subpopulation separately and compare the results to the population results, LOI concerns are well explained by differential SIPD across subpopulations. IPD and DIF in a single form were unable to provide explanations for the LOI results. With respect to dimensionality, the focal subpopulation fit indices were often noticeably different from the reference subpopulation fit indices, which indicates LOI in dimensionality. These subpopulation fit differences were best explained by the SIPD condition. With respect to scaling coefficients and ability estimates, the clearest explanation for LOI was the degree to which one subpopulation's item parameters drifted in a different manner than the other subpopulation's item parameters. Hence, the broad conclusion is that when subpopulation-level item parameters drift in different ways across subpopulations, measurement invariance is compromised and it is manifested in an unequal relative dimensionality of anchor tests, an incomparability of scaling coefficients across subpopulations, and a lack of equity in scaled ability estimates. More items with differential SIPD results in a larger magnitude of LOI, whereas sample size and mean differences were more varied in their relationships to the outcomes studied. These psychometric consequences are unexplained by the usual within-form DIF and IPD investigations performed in operational practice, which miss the important invariance concern caused by differential SIPD.

Practitioners often examine for DIF and IPD because researchers have discussed the connection between these item parameter invariance concerns and measurement validity (e.g., Han et al., 2012; Kane, 2006). This study demonstrated that these typical methodologies for examining invariance in item parameters fail to capture item parameter drift that is specific to subpopulations of examinees, and therefore fail to explain additional measurement invariance problems that compromise critical processes and outcomes of test data calibration, scaling, and reporting. While a previous study found that certain types of across-form DIF can result in a lack of equating invariance when using item response theory true score equating procedures (Huggins, 2014), the current study more formally defined this issue as differential SIPD and

then demonstrated the psychometric mechanisms by which the final outcome of equating invariance can be compromised. Specifically, when differential SIPD is present, dimensionality structure of the anchor test varies across subpopulations, which results in subpopulation-sensitive scaling relationships between the two administrations of the anchor test, which results in a lack of invariance in ability estimates. Any subsequent equating or scaling of these ability estimates will suffer from the same problems with invariance of scores across subpopulations. Yang and Gao (2008) called for studies that can further our understanding of the relationships between the same construct property of equating, the population invariance property of equating, and multidimensionality. This study contributes to that need. Differential SIPD results in differential anchor test dimensionality across groups (i.e., relative violation of unidimensionality in at least one of the groups), which is an indication of differential constructs being assessed across groups, the effects of which include a lack of invariance in scaling and equating outputs.

In the climate of high-stakes decisions based on subpopulation disaggregated test scores, it is critical that we understand all the ways in which subpopulations can interact differently with test items over time, and how those differences impact the scores reported from those tests. IPD and within-form DIF cannot always explain the subpopulation differences we may observe, making it difficult to identify the source of some LOI problems and attempt to correct them. If differential SIPD exists, it is a good explanation for some measurement invariance concerns, and knowing this can assist in the search for curriculum emphasis and/or opportunity to learn explanations for the observed problems. Even if a substantive reason for the presence of differential SIPD cannot be determined, knowing that one should estimate its presence and attempt to revise or remove items from anchor sets that demonstrate its presence can help control for additional LOI problems that arise. It is strongly recommended that practitioners examine for SIPD that differs across subpopulations.

Limitations and Future Directions

A first limitation of this study is that only b parameters were subjected to SIPD, but other parameters can also drift. Examining a parameter SIPD may be particularly important as it most likely has an impact on the A scaling coefficient, which was not impacted in this study and therefore not explored. Manipulating only b parameters and using only one scaling method helped to contain the myriad results of this study, but that is not to say that those things are not worthy of exploration in other studies. Second, the operational test that was mimicked in this study had less than 20% of the items serving as anchor items in a given form. It may be that the type and degree of SIPD in this study has less of an effect on the outcomes when the anchor set is larger. Future studies with larger anchor sets may be of use. Finally, the use of a simulation allowed for isolation of the study factors in order to understand causal relationships between those factors and the outcomes. However, research on SIPD in operational tests is needed to understand the ways in which it is commonly (or uncommonly)

manifested in item responses and estimated item parameters, as well as how those SIPD manifestations relate to measurement invariance concerns in practice. Last, for practitioners to utilize this study's results in operational test data, methods for detecting statistically significant differential SIPD are needed.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1996). Children in motion: School transfers and elementary school achievement. *Journal of Educational Research, 90*, 3-12.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement, 75*, 185-207.
- Battaaz, M. (2014). equateIRT: Direct, chain and average equating coefficients with standard errors using IRT methods. R package version 1.2. Retrieved from <http://CRAN.R-project.org/package=equateIRT>.
- Baxter, S. D., Royer, J. A., Hardin, J. W., Guinn, C. H., & Devlin, C. M. (2011). The relationship of school absenteeism with body mass index, academic achievement, and socioeconomic status among fourth-grade children. *Journal of School Health, 81*, 417-423.
- 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. 397-460). Reading, MA: Addison-Wesley.
- Bock, R. D., Muraki, E., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.
- Chen, Q. (2013). *Remove or keep: Linking items showing item parameter drift* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3560290)
- DeMars, C. (2012). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*, 265-300.
- Deng, H., & Melican, G. (2009, April). *An investigation of scale drift in computer adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

- 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., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT® mathematics test data across several administrations* (ETS Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service.
- Education Week. (2015). The Every Student Succeeds Act: Explained. *Education Week*, 35(14), 17.
- Giordano, C., Subhiyah, R., & Hess, B. (2005, April). *An analysis of item exposure and item parameter drift on a take-home recertification exam*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. Retrieved from <http://eric.ed.gov/?id=ED497708>.
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20, 369-377.
- 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, 97-117.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement*, 74, 627-658.
- Human Resources Research Organization, & Harcourt Assessment. (2006). *Reading and mathematics: Technical report for 2005 FCAT test administrations*. San Antonio, TX: Author. Retrieved online from <http://fcat.fldoe.org/fcatpub5.asp>.
- Joint Committee on Testing Practices. (2004). *Code for fair testing practices in education*. Washington, DC: Author. Retrieved from <http://www.apa.org/science/programs/testing/fair-code.aspx>.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Karger, J., & Hitchcock, C. (2003). *Access to the general curriculum for students with disabilities: A brief for parents and teachers*. Wakefield, MA: National Center on Accessible Instructional Materials. Retrieved from http://aim.cast.org/learn/historyarchive/backgroundpapers/brief_parents_teachers#.VRLwQOGnRod.
- Klein, A. (2015, December 10). President signs ESEA Rewrite, giving states, districts bigger say on policy. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2015/12/president_barack_obama_signs_e.html?cmp=enl-enl-eu-news1.
- Krause, J. (2012). *Assessment of item parameter drift of known items in a university placement exam* (Master's thesis, Arizona State University). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 963541958).
- Li, X., & Wang, W. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, 52, 28-54.
- Mislevy, R. J. (1982, March). *Five steps toward controlling item parameter drift*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*. 7th ed. Los Angeles, CA: Muthén & Muthén.
- No Child Left Behind: State Plans, 20 U.S.C. § 1111 (2001).
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). New York, NY: Elsevier.
- R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>.
- Rennie Center of Education Research and Policy. (2013). *Research meets policy: Opportunity to learn science?* Boston, MA: Author.
- Rossell, C. H. (2003). *Policy matters in teaching English language learners: New York and California*. New York, NY: ERIC Clearinghouse on Urban Education.
- Rupp, A. A., & Zumbo, B. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66, 63-84.
- Snyder, C. W., Mereku, K. D., Amedahe, F. K., Etsey, K., & Adu, J. (2013). National assessment and opportunity to learn in education reform. *International Journal of Educational Reform*, 22, 76-102.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- 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 ProQuest Dissertations and Theses database. (UMI No. 814814255)
- Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement*, 29, 201-211.
- U.S. Department of Education (2009). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author. Retrieved from <http://www2.ed.gov/admins/lead/account/peerreview/index.html>
- U.S. Senate (2015). *The Every Child Achieves Act of 2015* (draft bill). Washington, DC: Author. Retrieved from <http://blogs.edweek.org/edweek/campaign-k-12/Senate%20Bill.pdf>.
- Ujifusa, A. (2014, January 14). All 50 states now using testing or student performance for school accountability. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/state_edwatch/2014/01/all_50_states_now_using_testing_or_student_performance_for_school_accountability.html
- Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education*, 27, 214-231.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77-87.
- Yang, W. L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement*, 32, 45-61.
- Ye, M., & Xin, T. (2014). Effects of item parameter drift on vertical scaling with the nonequivalent groups with anchor test (NEAT) design. *Educational and Psychological Measurement*, 74, 227-235.

- Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32*, 62-80.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International.