Measuring the Power of Learning.®

Research Report
ETS RR–17-60

An Empirical Investigation of the Potential Impact of Item Misfit on Test Scores

Sooyeon Kim

Frederic Robin

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

An Empirical Investigation of the Potential Impact of Item Misfit on Test Scores

Sooyeon Kim & Frederic Robin

Educational Testing Service, Princeton, NJ

In this study, we examined the potential impact of item misfit on the reported scores of an admission test from the subpopulation invariance perspective. The target population of the test consisted of 3 major subgroups with different geographic regions. We used the logistic regression function to estimate item parameters of the operational items based on the empirical data accumulated over 3 years. A new set of item parameter estimates derived using the data from each subgroup separately was compared to the original (i.e., operational) item parameter estimates to assess the degree of item misfit due to subgroup memberships. Using the new set of item parameter estimates for each subgroup, we also updated the conversion tables, which were derived from the original item parameter estimates, and compared them to their original conversions to determine whether score invariance was achieved at the scaled score level. Score invariance was not absolutely achieved. Even so, the magnitude of reported score differences (systematic error or bias) caused by subgroup dependence was still smaller than the standard error of measurement (random error) of the test. This study suggests a practical remedy for enhancing the level of score invariance of the test.

Keywords Score invariance; multistage test; item DIF; item calibration and linking

doi:10.1002/ets2.12190

When test forms are assembled using well-established content and statistical specifications, the relative difficulties of different versions of a test will likely change as a function of score level in the same manner across subpopulations; thus, the different versions are related to each other in the same way across the subpopulations. If the relative difficulties of different forms interact with group membership, or if an interaction emerges among score level, difficulty, and group, score invariance is not achieved across the subpopulations. Under this circumstance, test takers having the same score on one scale may not have the same level of the proficiency being measured by the test because the score may depend on their group membership. This situation results in an advantage (or disadvantage) for one or more subpopulations, and hence, a lack of invariance in scores is a concern for fairness and equity in assessment (Dorans, 2004).

Subpopulation invariance (interchangeably with score invariance) is a concern for fairness and equity at the reported (equated or scaled) test score level. In practice, subpopulation invariance may not hold due to various reasons. A lack of equating invariance is often associated with a differential difficulty within the data that can be manifested in several ways, such as differential mean difficulties of items or differential rank ordering of anchor items in difficulty (Cook & Petersen, 1987; Dorans, 2004; Dorans & Holland, 2000; Holland & Dorans, 2006; von Davier & Wilson, 2008). Changes in item parameters can be an indirect source of subgroup dependence. Item parameter changes could occur due to changes in curriculum or frequent exposure of items because the items become easier or more difficult for the entire population. Such changes can threaten the validity of test scores by introducing trait-irrelevant differences on proficiency estimates. For example, an overexposed item becomes easier and less discriminating, causing errors in proficiency estimation using estimates of the original item parameters. Such change is further complicated in a situation in which the degree of change in item parameters varies depending upon the subgroup membership. Differences in parameter estimates across different subgroups are referred to as *differential item functioning* (DIF; Dorans & Holland, 1993).

DIF concerns fairness with respect to statistical bias at the item level (Camilli & Shepard, 1994), and it can be a typical source of item misfit.[1] DIF refers to the instance in which test takers of the same ability level have different estimated probabilities of success on a test item depending upon their group membership (Camilli & Shepard, 1994; Penfield & Camilli, 2007). A lack of invariance at the item level likely results in a lack of invariance at the reported test score level due to subgroup dependence. Dorans (2004) discussed how particular types of multidimensionality can affect both equating

Corresponding author: S. Kim, E-mail: skim@ets.org

invariance and DIF. von Davier and Wilson (2007) acknowledged that the presence of DIF in anchor items violates the unidimensionality assumption of IRT, which could pose issues for equating latent scores. Using simulated data, Huggins (2012) investigated the relationship between equating invariance and DIF in a systematic manner and showed that DIF manifested in the anchor items of an assessment can have an effect on population invariance of equating.

Often a lack of item invariance due to subgroup dependence is the main culprit in the lack of score invariance at the reported score level. As the literature indicates, a lack of invariance at the item level could become critical in situations in which a number of poorly fitting (misfit) items are used as an anchor set in score linking. The manifestation of item misfit in the anchor items may result in inaccurate and unfair scores for some subgroups of test takers. This issue will be compounded in situations in which not all subgroups are included in the linking sample. In reality, however, the particular group used to link two tests always affects the linking function, so that score invariance is never absolutely achieved across subpopulations. Instead, the question becomes whether score invariance holds closely enough that the equating function is not differentially affected across subpopulations (Dorans & Holland, 2000).

Purpose

The purpose of this study was to evaluate the potential impact of item misfit on test scores with respect to a lack of subpopulation invariance on equating. Some level of item misfit is unavoidable, particularly when testing across subpopulations with very diverse educational experiences and primary languages. For this investigation, we chose a large-scale international examination on which test takers were very heterogeneous in terms of their self-reported best language, region, social and cultural backgrounds, and so on. Three subgroups were classified as a function of region in which test takers took the test and their self-reported best language. At the item level, we analyzed the empirical data obtained from the operational setting (often called *post-admin data*) to find out the extent to which item parameter estimates in an item bank (i.e., original estimates calibrated a priori) differ from the new set of estimates derived using the response data from each of the three subgroups. At the score level, we then updated the original (i.e., operational) conversions separately for each subgroup using the new set of item parameter estimates derived using the post-admin data. We investigated the subpopulation invariance of score linking by comparing the original conversions with the new ones. Although a few researchers have examined the relationship between equating invariance and DIF (Huggins, 2012; von Davier & Wilson, 2007), the nature of the effect of general misfit items on population invariance in equating has yet to be empirically examined, particularly under the adaptive testing framework. Findings of such a study have practical implications.

The Current Practice

The target test used in this study followed a two-stage multistage testing (MST) procedure, in which one adaptation to the test takers' ability levels took place (see Figure 1). As shown in Figure 1, a two-stage MST form includes four modules across two stages. At Stage 1 (often called routing), there is only one module (20 items); all test takers taking that form are tested with same set of items. At Stage 2, there are three modules (20 items in each): a low-difficulty module, a medium-difficulty module, and a high-difficulty module. Each module at Stage 2 concentrates on a particular level of difficulty to differentiate test takers' abilities within a certain range of proficiency after routing. The items a test taker receives at Stage 2 are determined by the test taker's performance on Stage 1. The term *path* can be used to mean a combination of modules that could possibly be presented to a test taker. As illustrated in Figure 1, there are three paths in the two-stage MST form, and each path consists of the first-stage module and one of the second-stage modules. For test security, the program administers more than 100 MST forms constructed through the automated multistage test assembly processes in accordance with the content and statistical specifications during a particular testing period.

Each MST form also includes one pretest module. The items in the pretest module are not scored at that administration, but they are used to assemble future operational MST forms once they are linked to the common scale. In the operational setting, about 300 operational items administered in the routing module (Stage 1) are used as equaters to transform the pretest items' parameter estimates onto the common scale through the test characteristic curve method (TCC; Stocking & Lord, 1983). The common items (equaters) are drawn from the item bank rather than a particular single old form. The current practice, common-item linking to a calibrated item bank, is a flexible design because it allows the common-item set to be chosen from many previous MST forms rather than from a single form. After the transformation is completed,

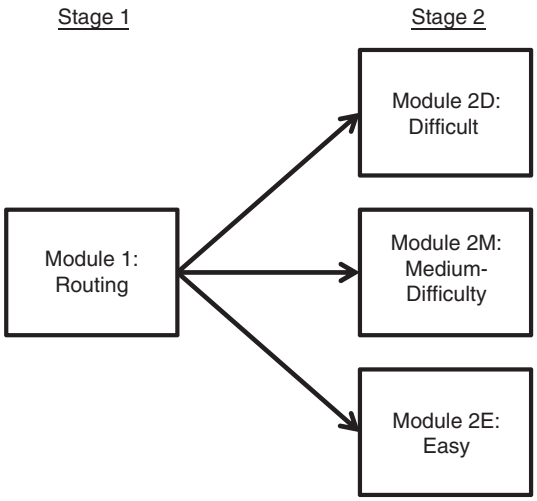

Figure 1 Schematic of a two-stage multistage test.

the pretest items have transformed item parameter estimates on the θ-scale that was previously established. These new items can be added to the IRT calibrated item bank, and the bank can be continuously expanded.

The test takers of the test are classified into three subgroups. Because those subgroups are heterogeneous in their performance level, the overall performance on each administration is heavily influenced by the proportion of each subgroup. To make the condition for item calibration and linking stable across administrations, a major subgroup has been used for item calibration and linking of new items in the operational setting. The item parameter estimates derived using the data from the major subgroup have been used to assemble the MST forms and to create the conversions from which the test takers' reported scores are derived. Under this circumstance, it is rather questionable whether the invariance property can hold for other subgroups excluded from the calibration and linking process. The present study was designed to assess this invariance using the real data of an admission test.

Method

Multistage Testing Forms

We chose 4,000+ MST forms administered in 2013. Therefore, 4,000+ MST forms were available for each of the three paths: *low*, *middle*, and *high*. Established before the administration of the new MST forms, all operational items have item difficulty and item discrimination values estimated using the two-parameter logistic model in the pretest item calibration and linking process. As soon as the new form operational items have been assembled into a form, their item parameter estimates can be used to compute a raw-to-scale score conversion table for the new form via IRT true score equating. Because raw-to-scale score conversions based on preequated item parameter estimates are available before the new forms are administered, each package per administration includes 100+ MST forms together with their conversion tables. Because all the test takers' proficiency estimates (i.e., $\widehat{\theta}$) and scaled scores are already on the common scale, they are comparable across not only different forms but also different paths.

MST forms can be scored in several ways. The current practice of the test is an inverse of TCC using number-correct scoring (i.e., summed scoring). According to Yen and Fitzpatrick (2006, p. 137), when tests include "30 or more items, the inverse of the TCC provides a very accurate MLE of ability for the 3PL model." Recently Kim, Moses, and Yoo (2015a, 2015b) and Kim and Moses (2016) compared the performance of seven IRT proficiency estimation methods under the two-stage MST design using simulated datasets. They showed that Bayesian estimators performed better than non-Bayesian ones, mainly at the two extreme score regions of the theta scale. Although the difference between item-pattern scoring and number-correct scoring was almost negligible in the portion of the score range where most test takers' scores were located, number-correct scoring produced more accurate ability estimates than item-pattern scoring for high-performing test takers. Even so, because the estimation results derived from different proficiency estimators were comparable across the theta region where most test takers would be located, the authors recommended the use of a

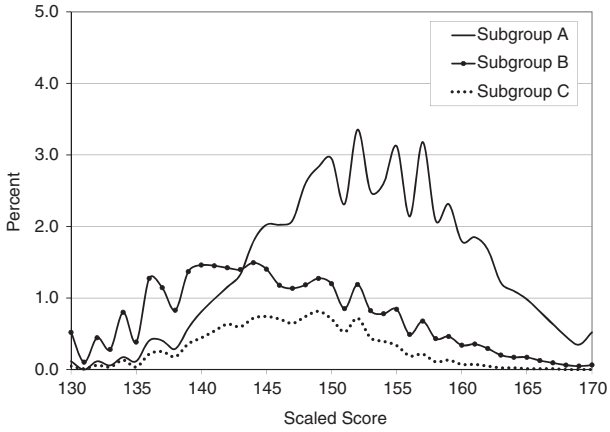

Figure 2 Relative frequency distribution of scaled scores in the three subgroups.

simpler method (e.g., non-Bayesian number-correct scoring) as a practical choice, particularly in the operational setting of the two-stage MST.

Data Analyses

The data used in this study were the test takers' actual responses to the operational items and their proficiency estimates accumulated over three testing years (2011–2014). The test takers were classified into one of the three subgroups based upon their stated best language, nationality, and testing region. The largest group (called Subgroup A hereafter), comprising approximately 60% of the test-taker pool, was used exclusively as the calibration and linking sample. The remaining test takers excluded from the linking sample were from two subgroups: Subgroup B (30%) and Subgroup C (10%). Figure 2 presents the relative frequency distributions of the scaled scores in the subgroups, given as percentages of the total group. Subgroup A is more proficient than the other two subgroups.

Logistic regression (PROC GENMOD with logit link function in SAS) was conducted for each item to obtain its new set of item parameter estimates (item difficulty and item discrimination) derived as a function of the test takers' responses to the item and their proficiency estimates that were derived from the original conversion.[2] In the logistic regression model, the dependent variable was the actual response on a particular item, scored as 0 (incorrect) or 1 (correct), and the independent variable was the proficiency estimates on the theta scale. A logit model that assumes binomial distribution of the probability of the event (0 or 1) was employed to estimate the intercept (β) and slope (α) parameters.

$$\text{logit}\{\Pr(Y=1|x)\} = \log\left\{\frac{\Pr(Y=1|x)}{1-\Pr(Y=1|x)}\right\} = \beta + \alpha x, \tag{1}$$

where Y is a response on a particular item (e.g., $Y=1$ if the answer is correct), and x is an explanatory variable, which was the proficiency estimate in the theta scale (θ) in this study. The slope (α) and intercept (β) parameters derived from the logistic regression model can be transformed to the IRT a (item discrimination) and b (item difficulty) parameters, as shown in Equation (2). See Hambleton, Swaminathan, and Rogers (1991, pp. 20–21).

$$a = \frac{\alpha}{D}, \text{ and } b = \frac{-\beta}{Da}, \quad \text{where } D = 1.702. \tag{2}$$

To ensure the stability of item parameter estimates, only items administered to at least 300 test takers per group were included for the analyses. Accordingly, all (strictly speaking, high-volume) items have two sets of estimates; one set comprised the original estimates derived from the operational setting (called "original" hereafter), and the other set comprised new estimates derived using the test takers' post-admin data (called "new" hereafter). For each item, logistic regression was repeated four times to obtain new item parameter estimates for the total group and each of three subgroups separately. For most operational items, four sets of new item estimates were available: (a) total, (b) Subgroup A, (c) Subgroup B, and (d) Subgroup C.

In the actual operational setting, the conversions for all paths of the MST forms were created using the 2PL IRT model-based item parameters that were estimated using the data from Subgroup A through the IRT true score equating procedure. Because those estimates were used operationally to report the score to the test takers, we called them "original" item parameter estimates. The same format of conversions was obtained for all the MST forms by replacing the original estimates with the new ones associated with the total and each subgroup. The new estimates are the ones derived from the logistic regression model applied to the empirical data. The new conversions were then compared to the original (operational) ones to determine whether the resulting conversions from each subgroup would yield scores comparable to the original conversions. If the population invariance property holds, then the resulting conversions from all three subgroups should be similar to the original conversions, leading to trivial differences in reported scores.[3]

Several thousand items were used to assemble the 4,000+ MST forms administered in 2013. Item parameter estimates of those items were updated using the post-admin data from the total group, Subgroup A, or Subgroup B, and thus about 4,000+ conversions per path were compared for each of those groups. However, this was not the case for Subgroup C. The new set of estimates was not available for all items due to a lack of data. Consequently, MST forms in which most items (35+ out of 40) were updated using the new estimates were selected for the comparison between the original and new conversions. The numbers of available MST forms after screening were 376, 1,064, and 1,030, from the high, middle, and low paths, respectively.

At the item level, we also compared the new and original item parameter estimates for each subgroup separately using two deviance measures commonly used to screen out misfit items from the equater set in the operational setting.[4] One is the unweighted maximum difference (UwMaxDiff), and the other is the weighted root mean square error (WRMSE). As the name indicates, UwMaxDiff is the maximum value among the differences between the new *item characteristic curve* (ICC) and original ICC. The WRMSE value indicates the root mean square of the sum of the differences between the two ICCs. The abilities used for the weighted comparison of the ICCs are from a normal distribution. Both measures were designed to detect any noticeable difference between two ICCs in terms of magnitude and pattern. The current practice designates any item whose UwMaxDiff is greater than 0.125 or WRMSE is greater than 0.1 as a misfit item. We used the same criterion, which is rather strict, in this study.

Results

Item-Level Analysis

Several thousand items were administered during the 3 years, and new parameter estimates for those items were obtained for each group separately using the logistic function. Table 1 presents the descriptive statistics of both the new estimates derived from the post-admin data and the original estimates. In Figure 3, the four plots in the first column, with the *x*-axis indicating original estimates and the *y*-axis indicating new estimates, present the relationships between the two sets of item discrimination estimates for total and Subgroups A, B, and C, respectively. The four plots in the second column present the same type of information for item difficulty.

The relationship between the original and new parameter estimates was generally stronger in item difficulty than in item discrimination across all groups. Because the operational samples used to produce the original item parameter estimates were essentially the same as Subgroup A, the means and *SD*s of estimates in Subgroup A were very similar to the original values of the discrimination (*a*) and difficulty (*b*) estimates, leading to very high correlations. As displayed in Figure 3, most items were evenly spread out along the diagonal line. The correlations (see Table 1) between the original and new estimates were close to unity, particularly for the difficulty parameter. Conversely, the relationship between the original and new estimates was very weak in Subgroup C, indicating substantial misfit at the item level, as can be seen by the estimates being broadly spread out along the diagonal line. There was more spread for the discrimination estimates, as would be expected, because those parameters are generally less precisely estimated than the difficulties. The discrimination estimates (*a*) of many items substantially increased compared to their original values, particularly in Subgroup C. In Subgroup B, the relationship between the two sets of estimates was not as strong as in Subgroup A but not as weak as in Subgroup C. Its overall relationship was rather moderate, as the graphical plots indicate.

Table 2 presents the summary statistics of the differences between the new and original estimates for the total and each subgroup. As presented in the upper part of Table 2, the means and *SD*s of the two difference values were generally small for total and Subgroup A. Both Subgroups B and C yielded large *SD*s, however, indicating more variation across

Table 1 Descriptive Statistics of Item Parameter Estimates for Total and Subgroups

Item parameter		Original	Total	Subgroup A	Subgroup B	Subgroup C
a	Mean	0.76 (0.78)	0.75	0.77	0.72	0.85
(Discrimination)	*SD*	0.27 (0.27)	0.27	0.27	0.27	0.43
b	Mean	−0.34 (−0.31)	−0.29	−0.32	−0.26	−0.20
(Difficulty)	*SD*	0.94 (0.92)	0.90	0.96	0.92	1.01
Correlation with original *a* estimate		—	0.85	0.93	0.75	0.37
Correlation with original *b* estimate		—	0.97	0.99	0.87	0.70

Notes. Due to limited sample size, about 20% of items did not have a new set of estimates in Subgroup C. For the comparison with Subgroup C, the means and *SD*s of the original estimates were recalculated after excluding those 20% items; those values are presented in the parentheses under the original column.

Table 2 Summary of Deviance Measures

Deviance	Group	Mean	*SD*	Min	Max
Difference *a*	Total	−0.01	0.15	−0.76	0.62
	Subgroup A	0.01	0.10	−0.44	0.55
	Subgroup B	−0.04	0.19	−1.07	0.78
	Subgroup C	0.07	0.42	−1.27	2.92
Difference *b*	Total	0.04	0.24	−1.28	2.99
	Subgroup A	0.01	0.14	−1.12	1.17
	Subgroup B	0.07	0.47	−2.84	7.15
	Subgroup C	0.11	0.76	−2.96	2.91
UwMaxDiff	Total	0.07	0.05	0.00	0.50
	Subgroup A	0.04	0.03	0.00	0.38
	Subgroup B	0.12	0.07	0.00	0.55
	Subgroup C	0.21	0.12	0.00	0.79
WRMSE	Total	0.04	0.03	0.00	0.21
	Subgroup A	0.02	0.02	0.00	0.13
	Subgroup B	0.08	0.05	0.00	0.37
	Subgroup C	0.14	0.09	0.00	0.57

the several thousand items. Again, this trend was much more salient in Subgroup C. In Figure 3, the four plots in the last column present the relationship between the discrimination difference and the difficulty difference. In those plots, the *x*-axis indicates the differences between new and original for the *a* estimates, and the *y*-axis indicates the differences between new and original for the *b* estimates. In general, the differences between two sets of estimates were scattered around the center area where the lines are crossed (i.e., $x = y = 0$). The differences in Subgroup A concentrated toward the center much more strongly than did those in Subgroups B and C. In Subgroup C, the differences dispersed widely over the area. This nondirectional trend is promising, because negative differences occurring in one set of items might cancel out positive differences occurring in another set of items.

The summary statistics of two deviance measures (UwMaxDiff and WRMSE) are presented in the lower part of Table 2. For both measures, the means of Subgroup C were much larger than the cutoff criteria of item misfit (e.g., UwMaxDiff > 0.125; WRMSE > 0.1), but the means of Subgroup B were slightly smaller than the cutoff criteria. As expected, the means of Subgroup A were very close to zero. Under the current practice, items whose UwMaxDiff is greater than 0.125 or whose WRMSE is greater than 0.1 are designated as misfit items. According to the current practice, the proportion of misfit items was about 10% in the total group, 2% in Subgroup A, 40% in Subgroup B, and 72% in Subgroup C. Because the current cutoff criteria are rather strict, many items were designated as misfit items in Subgroup C.[5] Despite the strict criteria, the proportion of misfit items was very small in Subgroup A, as expected.

Score-Level Analysis

Figure 4 depicts the averaged conditional scaled score differences between the new conversions and original conversions over the 4,000+ MST forms separately for each path and for each group, along with the 90% band denoted by dashed

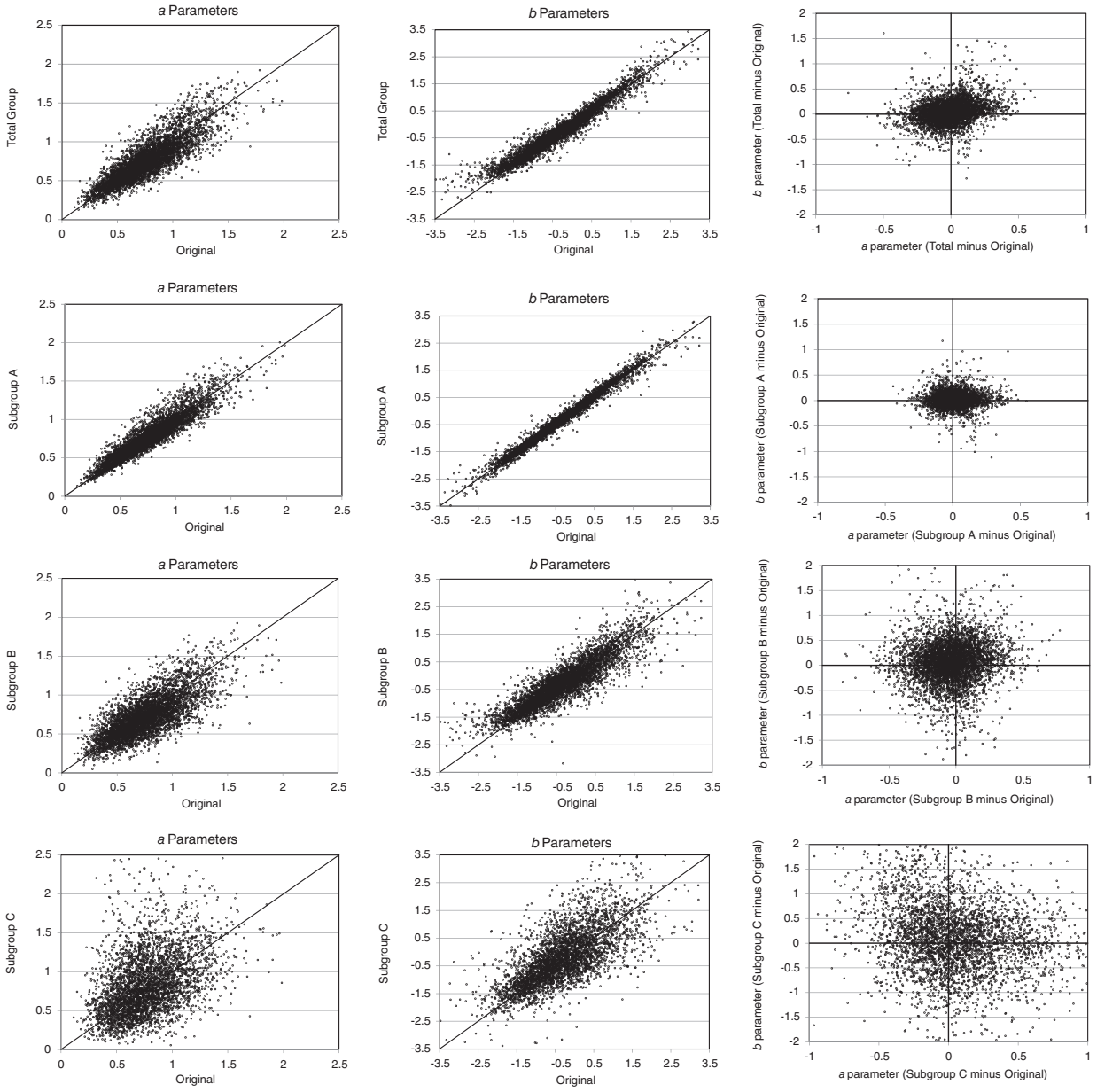

Figure 3 Item parameter estimates of *a* and *b* and difference plots for each of the total and the three subgroups.

lines. The 90% band indicates that 90% of difference values over the 4,000+ forms were located within this range.[6] The solid lines at ±1.00 were simply added to enhance the readability of the plots. By design, the possible range of scaled scores for each path varies due to the difference in form difficulty. For that reason, the *x*-axis indicating scaled scores was not identical across the three paths.

As expected, the averaged difference derived using Subgroup A was close to zero for most score points of all paths, indicating negligible systematic bias. As the narrow 90% band indicates, the variability of the difference across the several thousand MST forms was very small, indicating consistent small differences among the numerous forms. The same trend did not emerge for Subgroups B and C. Their 90% bands were much wider, indicating substantial variability among the MST forms. This trend was much more prominent for the high path. Some forms led to a difference as large as two score points, mainly at the extremes of the score scale. Across the entire score region, however, the maximum scaled score differences were generally smaller than the standard error of measurement of the test (e.g., 2.5). The systematic bias was a concern, particularly for the high path of Subgroups B and C. The averaged difference lines departed from the zero

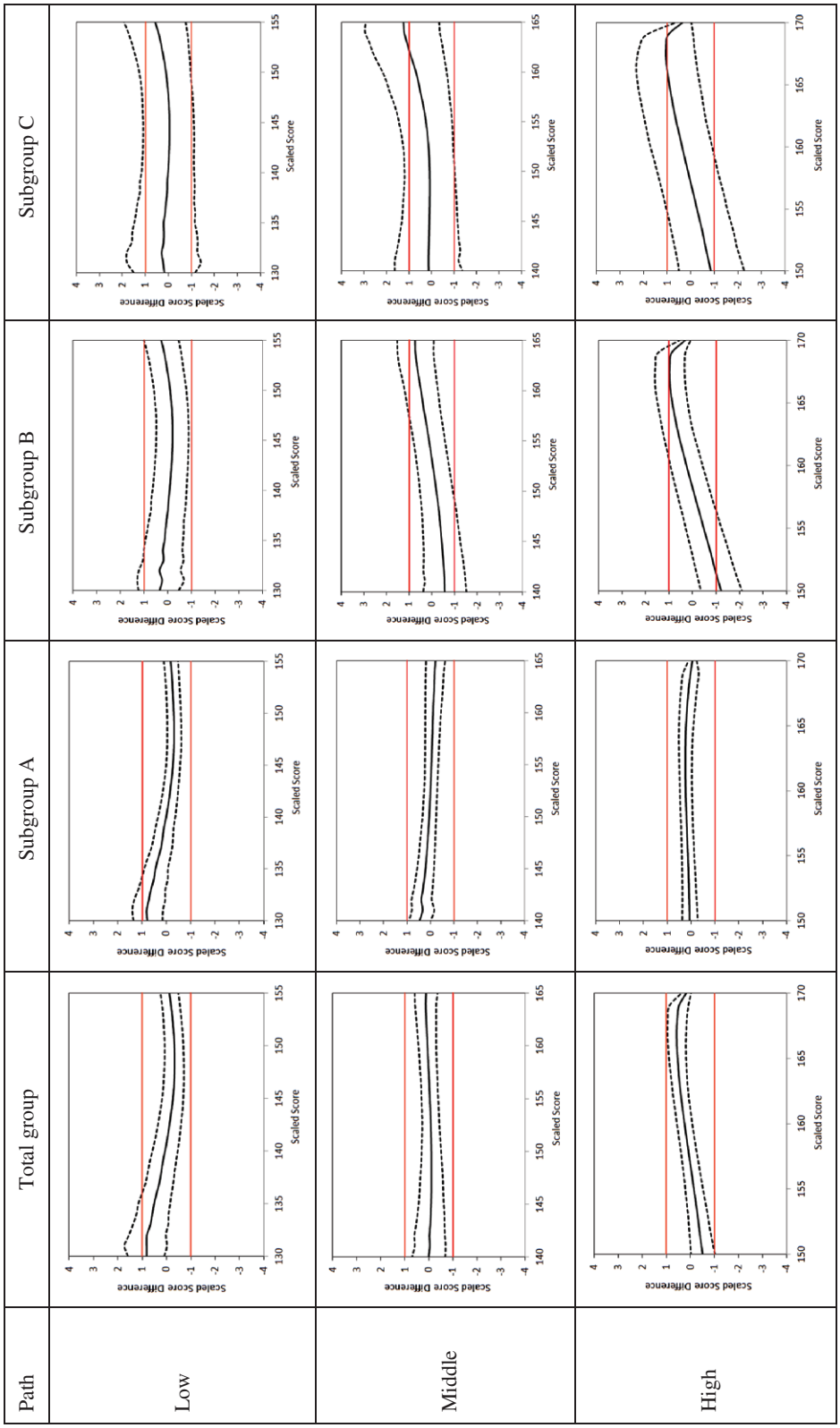

Figure 4 Scaled score differences between the new and original conversion for each path and of each subgroup. —— = Mean; ---- = 90% difference band, which ranged from the fifth percentile to the 95th percentile of the difference scores over the 4,000+ MST forms.

line, yielding negative differences at the lower end and positive differences at the upper end. The direction indicates that the new conversions cause the forms to appear easier at the lower end and more difficult at the upper end, as compared with the original conversions. The results from the total group were generally comparable to the results from Subgroup A, because Subgroup A comprised 60% of the total group. The difference pattern associated with the high path was somewhat different from the one associated with Subgroup A; however, the deviation from the zero line occurring in both Subgroups B and C was quite large.

Conclusions

Testing programs should be aware of possible statistical biases that may render test scores an unreliable measure for high-stakes decision making. According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), bias in measurement refers to the instance in which characteristics of a test result in different meanings of the scores earned by different subpopulations of test takers. A lack of subpopulation invariance (equating invariance) implies inconsistency in the relationship between the raw score scale and equated score scale as a function of group memberships (Dorans & Holland, 2000). If a lack of invariance occurring at the item level results in a lack of invariance at the reported score level, it is important to understand how a lack of item-level invariance arises and affects score equity across subpopulations.

In this study, comparisons of new and original parameter estimates in the score conversions revealed nontrivial differences in the reported scores, indicating a lack of subpopulation invariance (more likely score invariance, in this study). In most score equity assessment analyses, evidence of population dependence suggests the need to reevaluate test assembly specifications or linking methods. Here, the indicated problem is neither with the test specifications nor with the linking method. From the subpopulation invariance perspective, we focused solely on the problem related to a group of test takers exclusively used for item calibration and linking to the item bank score scale.

The current practice uses Subgroup A only as the sample for item calibration and linking. The significant difference between the conversions detected using the data from Subgroups B and C indicates that failure to include those subgroups in the operational calibration might bias the resulting linking. This tendency would increase if the proportions of those subgroups become larger over time. In reality, however, the use of Subgroup A only as a linking sample is defensible due to not only practical but also psychometric benefits. Use of stable linking samples over time is necessary to produce stable item parameter estimates and hence to maintain the quality and integrity of the item bank. The three subgroups of the test are heterogeneous in many aspects (e.g., proficiency, language), and the proportions in Subgroups B and C compared to Subgroup A are not constant across the administrations. According to a test form assembly plan of the test, item calibration and linking of new items has to be conducted separately for each administration. Under this circumstance, using all test takers for linking may add unwanted bias to the item bank due to the instability of linking samples. In addition, not all subgroups are equally vulnerable to test security violations, which can be a threat to the validity of test scores. In the past, test security concerns such as cheating were constantly raised with Subgroup C. Inclusion of other subgroups, particularly Subgroup C, in the linking sample may exacerbate the situation by adding different sources of bias to test scores. The continuation of the current practice (using Subgroup A only) will be a safe choice for this test to maintain the stability of the reporting scale over time.

Even so, it is worth noting that there are limitations in generalizing the findings of the current study and the choice of linking sample to other testing programs. The subpopulation invariance properties depend on the definition of subgroups and the characteristics of the proficiency being measured. Invariance of a certain subpopulation on a certain test cannot be generalized to other subpopulations or other tests. In addition, a choice of linking sample could be different depending on the score region that is of most interest (the entire score scale, or a cut score region).

We suggest a different remedy for enhancing the level of subpopulation invariance of the test. Often a lack of item invariance is a main culprit in the lack of linking invariance at the reported score level. It is useful to conduct statistical checks to discover which operational items function differently across subgroups after adjusting for subgroup members' differences in ability. As shown at the item level, many operational items in the bank yielded nontrivial differences compared to the item parameter estimates derived using Subgroup C. The apparent difficulty of some items decreased or increased dependent upon subgroup membership. The same trend was noted with the discrimination parameters. It is unrealistic to expect that the difficulty of all forms of a test can be eventually balanced by using negatively biased items to cancel the effects of positively biased items in each form. Excluding problematic misfit items from the item bank would be an alternative

method to reduce statistical bias in the resulting linking function. If a large proportion of items function differently across the subgroups, however, the removal of those items from the item bank can be another threat to the validity of the test due to the reduction in the content coverage. Furthermore, this remedy will lead to a practical challenge, such that item writers must employ greater specificity in item development, resulting in an overall cost increase. From the practical perspective, excluding misfit items from the equater set for linking will be a better choice than excluding them permanently from the item bank. In this situation, misfit items would be eligible for scoring, as would other operational items, but they would not be eligible for use as equaters, thus ensuring the fairness of linking new items to the item bank. Additional empirical investigation to detect interaction effects between item type and misfit would be worthwhile to enhance the validity of test score uses.

Using real data on an admission test, we examined score invariance at the item level as well as at the reported score level in order to assess the quality of score linking. This topic provides another avenue for future investigation to determine adequate criteria to flag misfit items in practice. The score comparability will be a concern in a situation in which many items with large misfit exist in the item bank, whereas the automatic test form assembly will be challenged in a situation in which the item resources are limited due to the rejection of many items, even with moderate misfit. In order to maintain the quality and integrity of the item bank, an appropriate compromise needs to be made between retaining misfitting items and rejecting misfitting items, depending upon various factors. The decision on what criteria are appropriate to flag misfitting items and to remove them permanently from the bank depends on several factors, such as test specifications (content and statistical), test score use (e.g., admission or certification), test design (i.e., linear, MST, or Computerized Adaptive Test [CAT]), or item bank size. Literature on this matter is lacking. To offer some practical guidelines to practitioners, simulation studies would be recommended to examine the extent to which misfit items manifest their effects through the automatic test form assembly process. Not only the degree of misfit but also the proportion of misfit items in the bank can be manipulated as study conditions.

Notes

1 We used *item misfit* as a broader concept than *DIF* throughout the paper.
2 The "person by item" data matrix accumulated over three testing years was not only extremely large but also very sparse due to the MST adaptive nature as well as its complicated form assembly design (e.g., 100+ forms per administration). Under this circumstance, use of any conventional IRT software was not feasible. Thus, we employed the logistic regression model to the empirical data in order to derive IRT *a* and *b* parameter estimates. Note that the new item parameter estimates derived from the logistic regression model would be an approximation of the IRT–logistic model-based item parameter estimates.
3 It is worth noting that the comparisons conducted in this study were rather different from the conventional comparisons for assessing the subpopulation invariance property. Because Subgroup A has been used only for the operational item calibration and linking, the score conversions created using the item parameter estimates derived from the total group were not available in the actual operational setting. Following the conventional subpopulation invariance approach, however, we compared each subgroup conversion derived using the new estimates associated with each subgroup to the total group conversion derived using the new estimates of the total group. We can provide the result upon request.
4 Some relevant information related to those indices can be found in the *GENASYS Statistical Manual* (IRT: TRANCOMP COMPARE). We can provide it upon request.
5 We compared the item parameter estimates derived from the logistic regression model to those derived from IRT calibration. Some difference between them may be due to the use of a different estimation model.
6 The 90% band ranged from the fifth percentile of the difference score to the 95th percentile of the difference score over the 4,000+ MST forms.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

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

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: Erlbaum.

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

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

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.

Huggins, A. C. (2012). The effect of differential item functioning on population invariance of item response theory true score equating. *Open Access Dissertations*. Paper 724. Retrieved from http://scholarlyrepository.miami.edu/oa_dissertations/724

Kim, S., & Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (GRE Board Research Report No. 16-03, ETS Research Report No. RR-16-22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12111

Kim, S., Moses, T., & Yoo, H. (2015a). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*, 70–79.

Kim, S., Moses, T., & Yoo, H. (2015b). *Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing* (Research Report No. RR-15-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12057

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics*: Vol. *26*. Psychometrics (pp. 125–167). New York, NY: Elsevier.

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

von Davier, A. A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement*, *67*, 940–957.

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

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Suggested Citation:

Kim, S., & Frederic, R. (2017). *An empirical investigation of the potential impact of item misfit on test scores* (Research Report No. RR-17-60). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12190

Action Editor: James Carlson

Reviewers: Hongwen Guo and Ying Lu