Measuring the Power of Learning.®

Research Memorandum
ETS RM–19-07

Psychometric Evidence to Assess Expanded Test Use

Sooyeon Kim

Seunghee (Sam) Chung

June 2019

ETS Research Memorandum Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Consultant

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Ariela Katz
Proofreader

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 Memorandum 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 Memorandum 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.

Psychometric Evidence to Assess Expanded Test Use

Sooyeon Kim
Educational Testing Service, Princeton, New Jersey

Seunghee (Sam) Chung
AICPA, Ewing, New Jersey

June 2019

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

Suggested citation: Kim, S., & Chung, S. (2019). *Psychometric evidence to assess expanded test use* (Research Memorandum No. RM-19-07). Princeton, NJ: Educational Testing Service.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/

To obtain a copy of an ETS research report, please visit http://www.ets.org/research/contact.html

Action Editor: Marna Golub-Smith

Reviewers: Hongwen Guo and Wen-Ling Yang

Copyright © 2019 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING are registered trademarks of Educational Testing Service (ETS). All other trademarks are the property of their respective owners.

Abstract

Testing programs are often called upon to use their tests for multiple purposes. Alternate uses may include administering the test to groups different from the target population. Whenever the use of a test is expanded by administering it in other than the target population, fairness becomes a major concern. This study illustrates how the evaluation of two psychometric concepts—differential item functioning and linking invariance—can provide useful empirical evidence to assess this extended test use. Data from a large-scale language examination on which test takers were heterogeneous not only in their backgrounds (e.g., language, geographic region) but also in their performance levels were chosen as an example.

Key words: repurposing test, linking invariance, differential item functioning, subpopulation invariance

Testing programs are often called upon to use their tests for multiple purposes. Alternate uses may include administering the test to populations outside of the target population. For example, stakeholders may want to use a test designed for a domestic population internationally or a test (e.g., a language test) designed for a specific geographic region globally. When it comes to educational testing, expanding test use in this way without the same rigor of validity evidence as that provided for its original use is problematic because test scores have significant consequences for test takers and score users (Wendler & Powers, 2009). Any expansion of a test beyond its originally intended use must maintain fairness to all examinees (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Wendler and Powers (2009) used a simple definition of repurposing, namely, "using a test either for test takers or for purposes that are different from those for which the test was originally developed" (pp. 1–2). They illustrated several sets of standards and guidelines that can be useful for deciding how to repurpose a test and interpret its scores appropriately. Among these standards, the authors viewed validity—the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests—as the top priority. Even if the testing program made great efforts to support the test's original purpose, more validity research may be needed to support the clams that might be made about the test's outcomes in its new context and, furthermore, to make the repurposed test scores fair, meaningful, and defensible.

Whenever we test multiple populations with the same instrument, we need to be mindful of interactions of group membership with measured content dimensions, which could impede fair assessment of test-taker ability (Castellano & Kolen, 2016). Testing programs may gather substantial amounts of psychometric evidence to support their original claims about a test. Regardless of the strength of that evidence, given that test scores can carry significant consequences not only for test takers but also for score users, it is equally important to evaluate evidence related to new populations as well.

Dorans (2004) introduced the concept of *score equity assessment* to cover various methodologies and procedures aimed at providing evidence that a test score measures what it measures in the same way across different populations and subpopulations. He saw invariance as central to any discussion of fairness. He distinguished at least three types of invariance: item-

level invariance, test-level invariance, and predictive invariance. Any or all of these types of invariance could be challenged by using a test in other than its original target population. At the item level, we refer to lack of invariance as differential item functioning (DIF). At the test level, we call lack of invariance differential test functioning.

Dorans (2004) saw the notion of subpopulation invariance of equating functions as central to the assessment of score equity. In principle, when test forms are well constructed and appropriate for all test takers, the relative difficulties of different test forms will likely change as a function of score level in the same manner across subpopulations; that is, the relationship between the two test forms is invariant across subpopulations. In this case, the two test forms are equatable, and the resulting scores may be considered equivalent across test forms. If the relative difficulties of different forms interact with group membership, test takers having the same score across test forms may or may not have the same level of proficiency being measured by the test. The resulting scores cannot ensure fairness to test takers. Subpopulation invariance of linking functions, then, is a necessary condition for invariant test functioning across subpopulations.

Often a lack of item invariance due to subgroup dependence, or DIF, is a main culprit in the lack of scoring invariance at the reported score level. DIF concerns fairness with respect to statistical bias at the item level, and it refers to the instance in which examinees of the same ability level have different estimated probabilities of success on a test item, depending on their group memberships (Camilli & Shepard, 1994; Penfield & Camilli, 2007). Multiple sources (e.g., preknowledge, language translation, curriculum change, cultural difference) could lead to DIF, depending on the disparity among the various populations (Elosua & López-Jaúregui, 2007; Gierl & Khaliq, 2001). DIF analysis is a key component in the evaluation of the fairness and validity of educational tests (Zwick, 2012). As the literature has indicated, a lack of invariance at the item level could become critical in situations where items with DIF are used as anchors in score linking (Huggins, 2012). This issue will be compounded in situations where not all subpopulations are included in the linking sample due to test security concerns. Often testing programs use the item parameters derived from a particular subgroup to assemble the test forms and create the conversions from which the examinees' reported scores are derived (Kim & Robin, 2017).

In this study, we focus on repurposing by expansion of the testing population. Given that condition, the purpose of this study is to illustrate how to evaluate the appropriateness of

expanded test use exclusively from the psychometric perspective. We focus on two psychometric concepts, item DIF and score linking, which are considered key components in the evaluation of the fairness of educational tests. With both analyses, we want to determine if the test behaves differently in a new population of interest than it does in the current population. Such differences may signal the inappropriateness of the test for the new population. Data from a large-scale language examination on which test takers were heterogeneous not only in their backgrounds (e.g., language, geographic region) but also in their performance levels were chosen as an example. We illustrate via the example how the two psychometric concepts can be evaluated in such a way as to produce useful empirical evidence either to support or contraindicate the extended use of a test.

Method

Data

The test used in this study consists of two sections: Section 1 and Section 2. Each section measures a different skill using 100 multiple-choice items and has an independent reporting scale. The data used in this study were the responses to the items in each section. A total of 132,486 test takers took the test in a single administration. They were classified into one of two subgroups (subpopulations) based on their nationality or geographic region. Subgroup A comprised approximately 52% of the examinees, and Subgroup B comprised the remaining 48% of examinees. Table 1 presents the descriptive statistics associated with each subgroup for each section. Figure 1 depicts the relative frequency distributions of Section 1 raw scores in the two subgroups, given as percentages of the total group. Figure 2 depicts the same type of distributions for Section 2. Subgroup B outperformed Subgroup A on both sections. The standardized mean difference between Subgroup B and Subgroup A was 0.95 in Section 1 (i.e., almost 1 standard deviation apart on the raw score scale) and 0.59 in Section 2.

Table 1. Descriptive Statistics of Section Raw Scores for the Total and Each Subgroup

Group	*N*	*M*	*SD*	Percentile (5th–95th)	Percentile (1st–99th)
Section 1 Total	132,486	71.53	16.10	42–94	33–98
Section 1 Subgroup A	68,951	64.96	15.73	39–90	32–96
Section 1 Subgroup B	63,535	78.67	13.18	53–95	39–98
Section 2 Total	132,486	60.92	17.36	32–89	25–95
Section 2 Subgroup A	68,951	56.23	16.68	30–85	24–93
Section 2 Subgroup B	63,535	66.00	16.64	36–91	27–96

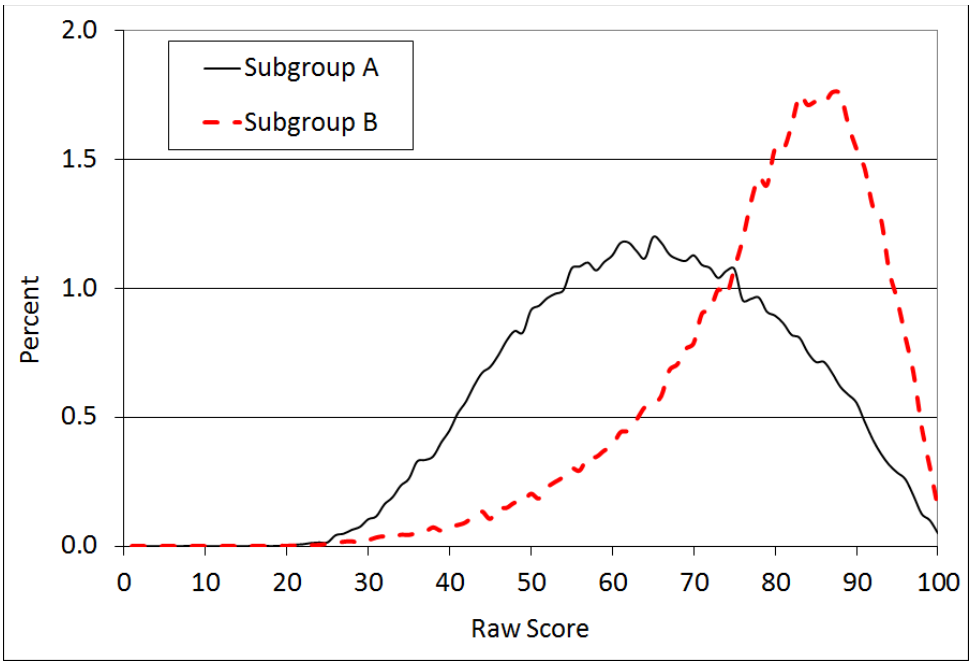

Figure 1. Relative frequency distribution of number-correct raw scores in the two subgroups: Section 1.

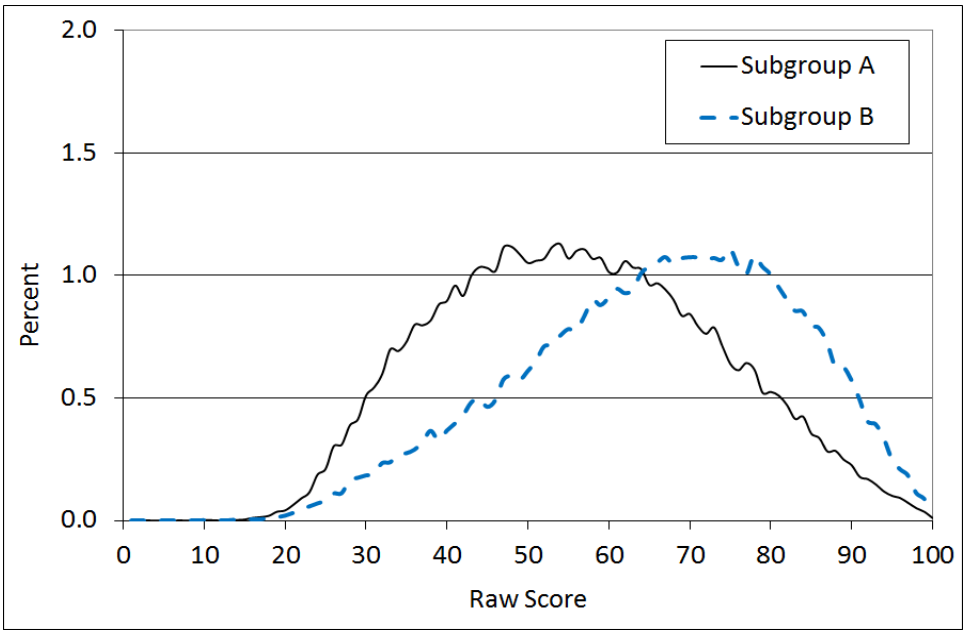

Figure 2. Relative frequency distribution of number-correct raw scores in the two subgroups: Section 2.

Procedure

DIF and subgroup linking invariance were used to evaluate the appropriateness of extended use at the item level and at the test score level, respectively. In the following analyses, Subgroup A was treated as an original target and reference group of the test, whereas Subgroup B was treated as a new test-taker group and focal group of the test for the sake of convenience.

Item-Level Analyses: Differential Item Functioning

For each item, DIF analysis compares the performance of members of the focal group with members of the reference group who have the same scores on a matching criterion to identify items that were particularly hard or particularly easy for a specified group of examinees. In this study, the total score on each section was used as the matching criterion.

The statistic used to indicate the size of the difference was the Mantel–Haenszel statistic, expressed in terms of the ETS delta scale (abbreviated as "MH D-DIF"). An MH D-DIF value of zero would indicate that reference and focal groups, matched on total score, performed similarly. An MH D-DIF value of 1.00 would indicate that the focal group (compared to the matched reference group) found the item to be 1 delta point easier, while an MH D-DIF of −1.00 would indicate that the focal group (compared to the matched reference group) found the item to be 1 delta point more difficult. Based on the results of the DIF analysis, the items are categorized into three classification levels (Educational Testing Service, 1988), where statistical significance is determined using $p < .05$: A = low DIF (absolute value of MH D-DIF less than 1 or not significantly different from 0); C = high DIF (absolute value of MH D-DIF at least 1.5 and significantly greater than 1); and B = neither A nor C. Each B and C DIF item can be further classified into two categories: B+/B− and C+/C−. Categories B+ and C+ include items that were easier for the focal group than for the matched reference group; Categories B− and C− indicate items that were more difficult for the focal group than for the matched reference group.

Many testing programs use criterion refinement procedures in conducting DIF analyses. Refinement is intended to improve the quality of the matching variable by removing items identified as having DIF (mainly C DIF) in a preliminary round of analysis. According to Zwick (2012), refinement of the matching criterion improves detection rates when DIF is primarily in one direction but can depress detection rates when DIF is balanced. She mentioned that refinement is advisable if nothing is known about the likely pattern of DIF. In this study, DIF analyses were repeated three times, in each section, via criterion refinement procedures.

Item Response Theory Item Calibration and True Score Equating: Score-Level Analysis

Using each subgroup's item responses, item calibration/linking and item response theory (IRT) true score equating were conducted to produce score conversions. For each section, item calibration/linking procedures were as follows:

1. Calibrate items based on Subgroup A's item responses under a 2-parameter logistic IRT model using flexMIRT (Cai, 2017).
2. Repeat Step 1 using Subgroup B's item responses.
3. Conduct the Stocking and Lord (1983) transformation using the computer program ST (Hanson & Zeng, 1995) to find the linear transformation function that places both sets of calibrated items on the same scale. At this step, use Category A items (from the third round of DIF analysis) as common items for linking.
4. Evaluate the transformation result from Step 3 based on two deviance measures: (a) the unweighted maximum difference (UwMaxDiff), which indicates the maximum value among the differences between item characteristic curves (ICC) of Subgroup A and ICCs of Subgroup B, and (b) the weighted root mean square error (WRMSE), which indicates the root mean square of the sum of the differences between the two ICCs.
5. Repeat Step 3 after removing items with an UwMaxDiff greater than 0.15 or a WRMSE greater than 0.1 from the common item set.
6. Rescale Subgroup B's item parameter estimates using the linear transformation function from Step 5 to put them on a common metric together with Subgroup A's parameter estimates.

The test score on each section is the sum of the number of correct answers on that section. So that scores will be equivalent across all test forms, this raw number correct is transformed to an equated number correct, that is, the number of correct responses on a reference form that is equivalent to a given number correct on the new form (assuming the same level of test-taker proficiency). IRT true score equating accomplishes this transformation (see Kolen & Brennan, 2004, pp. 191–194). Figure 3 is a schematic of IRT true score equating used in this study.

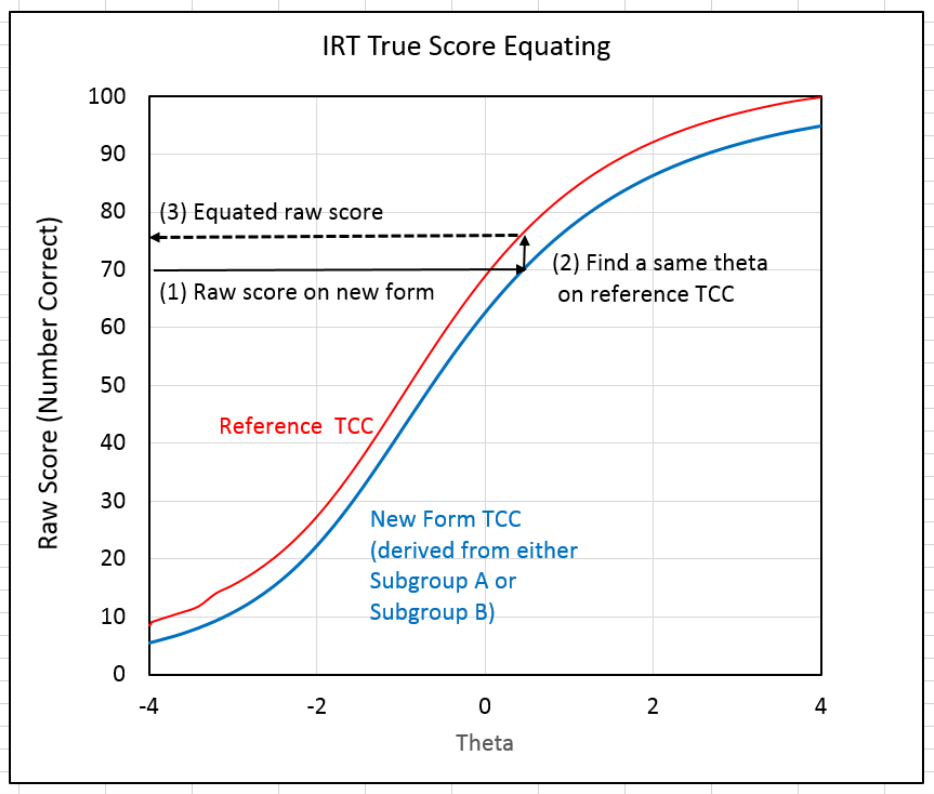

Figure 3. Schematic of item response theory true score equating.

In the context of IRT true score equating, the reference form is essentially a lookup table relating test scores to examinee proficiency level (theta). Thus a testing program might set the reference form in any of several ways. The reference form may be the first form given operationally, or it may be some hypothetical form with some ideal set of characteristics. For the sake of this illustration, we chose the latter option, creating a hypothetical form containing 100 items. Before IRT true score equating can be used, the item parameters for the new and reference forms need to be on the same scale. To simplify this illustration, we assumed that the parameter estimates for the new form were already on the reference form scale. For each section of the test, IRT true score equating procedures were as follows:

1. Obtain a test characteristic curve (TCC) based on the *a* and *b* parameters on a hypothetical 100-item reference form.[1]
2. Obtain a new form TCC using Subgroup A's item parameter estimates to define a one-to-one correspondence between the number-correct score on the test and examinee latent proficiency (i.e., theta).
3. Find a theta associated with each number-correct raw score on the new form TCC and then find the score on the reference form TCC associated with the same theta. The reference form equivalent score is essentially the same as the equated raw score.
4. Obtain a raw-to-equated raw score conversion from Step 3.
5. Repeat Steps 2–4 using Subgroup B's item parameter estimates to obtain Subgroup B's raw-to-equated raw score conversion.

Test Score Level: Subpopulation Linking Invariance

After true score equating, Subgroup B's score conversion was compared to Subgroup A's score conversion to determine whether the resulting conversion from Subgroup B would yield scores comparable to the conversion of Subgroup A. If the subpopulation linking invariance property holds, then the resulting conversions from both subgroups should be similar. The differences were compared to the *difference that matters* (DTM; Dorans & Feigenbaum, 1994), defined as half of a raw score interval, across the score region where most test takers were located.

The Subgroup A conversion was applied to every test taker of Subgroup B to obtain an equated raw score for each section. This score is the raw score the test taker would have received if the Subgroup A conversion had been used for all test takers. In addition, the Subgroup B conversion was applied to every test taker of Subgroup B to obtain his or her equated raw score for each section. This score is the raw score the test taker would have received if the Subgroup B conversion had been implemented in practice. The score differences caused by using different conversions (either the Subgroup A or the Subgroup B conversion) were calculated to assess how many test takers' scores would change. Based on those differences, we computed the percentage of test takers whose rounded equated scores were categorized as follows: no difference, a 1-point difference, a 2-point difference, and so on. In addition, the means and standard deviations were

computed based on the two sets of unrounded raw scores derived from the two subgroup conversions.

Results

Item-Level DIF Analysis

Tables 2 and 3 present the summary of DIF analyses in Sections 1 and 2, respectively. Under each section, the first column indicates the first round of DIF analysis using the sum of 100 item scores as a criterion score. The second column indicates the second round of DIF analysis using a refined criterion score, which is the sum of scores of items in Categories A and B from the first round. The third column indicates the third round of DIF analysis.[2] Table 4 presents the number of items that belong to each of the 5 × 5 category combinations determined by the first and third round results for Section 1. Table 5 presents the same type of information for Section 2. In the tables, the shaded cells indicate the items whose DIF categories remained the same over the two rounds. For Section 1, 96 items maintained the same DIF categories across Rounds 1 and 3. For Section 2, 90 items were consistently categorized. In both sections, the DIF results are very stable, regardless of refinement on the criterion score.

As shown in Tables 2, 3, and 4, about two thirds of items were designated as Category A and about 20% of items were designated as Category C in Section 1. Among the C DIF items, about two thirds favored Subgroup B (C+) over Subgroup A, whereas about two thirds of B DIF items favored Subgroup A (B−) over Subgroup B. The opposite direction of favoritism may balance out the impact of high and moderate DIF items. This trend did not appear in Section 2. As shown in Tables 2, 3, and 5, over 70% of items were designated as Category A in Section 2. The proportion of Category C items was lower than 10%, but most of them favored Subgroup A (C−) over Subgroup B. About 20% of Category B items did not clearly favor a particular subgroup. In practice, often, it is not easy for content experts to explain the reason why certain items perform differently as a function of test takers' group memberships.

Table 2. Summary of Differential Item Functioning (DIF) Analyses Between Subgroup A and Subgroup B in Section 1

DIF category	Criterion: Sum of 100 item scores	Criterion: Sum of 81 item scores	Criterion: Sum of 80 item scores
A	66	67	67
B+	6	5	5
B-	9	10	10
C+	11	12	12
C-	8	6	6

Note. $N = 100$. Categories B+ and C+ include items that were easier for the focal group (Subgroup B) than for the matched reference group (Subgroup A); Categories B- and C- indicate items that were more difficult for the focal group (Subgroup B) than for the matched reference group (Subgroup A).

Table 3. Summary of Differential Item Functioning (DIF) Analyses Between Subgroup A and Subgroup B in Section 2

DIF category	Criterion: Sum of 100 item scores	Criterion: Sum of 93 item scores	Criterion: Sum of 91 item scores
A	71	72	75
B+	13	9	8
B-	9	11	10
C+	2	1	1
C-	5	7	6

Note. $N = 100$. Categories B+ and C+ include items that were easier for the focal group (Subgroup B) than for the matched reference group (Subgroup A); Categories B- and C- indicate items that were more difficult for the focal group (Subgroup B) than for the matched reference group (Subgroup A).

Table 4. Differential Item Functioning (DIF) Category Patterns in Section 1

DIF category	3rd DIF A	3rd DIF B+	3rd DIF B−	3rd DIF C+	3rd DIF C−	Total
1st DIF A	66	0	0	0	0	66
1st DIF B+	0	5	0	1	0	6
1st DIF B−	1	0	8	0	0	9
1st DIF C+	0	0	0	11	0	11
1st DIF C−	0	0	2	0	6	8
Total	67	5	10	12	6	100

Note. Five shaded cells (which are, by row and column, A, A; B+, B+; B−, B−; C+, C+; C−, C−) indicate the items whose DIF categories remained the same over the two rounds. 1st category DIF criterion = sum of 100 item scores. 3rd category DIF criterion = sum of 80 item scores. DIF = differential item functioning. $N = 100$.

Table 5. Differential Item Functioning (DIF) Category Patterns in Section 2

DIF category	3rd DIF A	3rd DIF B+	3rd DIF B−	3rd DIF C+	3rd DIF C−	Total
1st DIF A	69	0	2	0	0	71
1st DIF B+	6	7	0	0	0	13
1st DIF B−	0	0	8	0	1	9
1st DIF C+	0	1	0	1	0	2
1st DIF C−	0	0	0	0	5	5
Total	75	8	10	1	6	100

Note. Five shaded cells (which are, by row and column, A, A; B+, B+; B−, B−; C+, C+; C−, C−) indicate the items whose DIF categories remained the same over the two rounds. 1st category DIF criterion = sum of 100 item scores. 3rd category DIF criterion = sum of 91 item scores. DIF = differential item functioning. $N = 100$.

Score-Level Analysis

For each section, separate subgroup calibrations were conducted. Category A items designated at the third round of DIF analysis (67 items in Section 1 and 75 items in Section 2) were used to determine Subgroup A's TCC as well as Subgroup B's TCC. The same items were used to find *the best linear transformation* (TBLT) of Subgroup B's item parameter estimates that minimizes the difference between the two TCCs. In Section 1, four items led to deviance values larger than the cutoff criteria (e.g., UwMaxDiff > 0.15 or WRMSE > 0.10) from the first TBLT. For the second TBLT, those misfitting items were removed from the common item set, and the remaining 63 items were used to derive the linear constants that minimize the difference between Subgroup A's TCC and Subgroup B's TCC. Furthermore, those linear transformation constants were applied to Subgroup B's item parameter estimates (from a separate calibration) to put them on the same scale together with Subgroup A's item parameter estimates. The same procedures were applied to Section 2. A total of 70 items, after removing 5 misfitting items from the common item set, were used to find the transformation constants.

Once the TBLT procedures were complete, IRT true score equating was conducted to find the extent to which the score conversion derived from Subgroup B differed from the score conversion derived from Subgroup A.[3] Because all C DIF items were assumed to have a single best answer, all 100 items were scored and used to produce the score conversions in each section.

In Figure 4, the solid red line depicts the conditional score differences between the Subgroup B conversion and the Subgroup A conversion at each raw score for Section 1. In the same figure, two dotted horizontal lines at ±0.5 indicate the DTM band, and two black vertical lines indicate the score range of the 1st (39) to 99th (98) percentiles in Subgroup B. In Section 1, the differences associated with raw scores from 8 to 55 were larger than the DTM of 0.5, and the largest difference between the two conversions was −1.6 at a raw score of 27. The differences over the raw score region higher than 55, where most test takers were located (see Table 1 and Figure 1), were smaller than the DTM.

Figure 5 presents the conditional score differences at the test score level for Section 2. In Section 2, the differences beyond the DTM band occurred over the raw score regions from 5 to 27 and from 40 to 83. The differences related to raw scores from 40 to 83 are problematic because many Subgroup B test takers' scores were within that range (see Table 1 and Figure 2). The largest difference was about 1.5 at a raw score of 61.

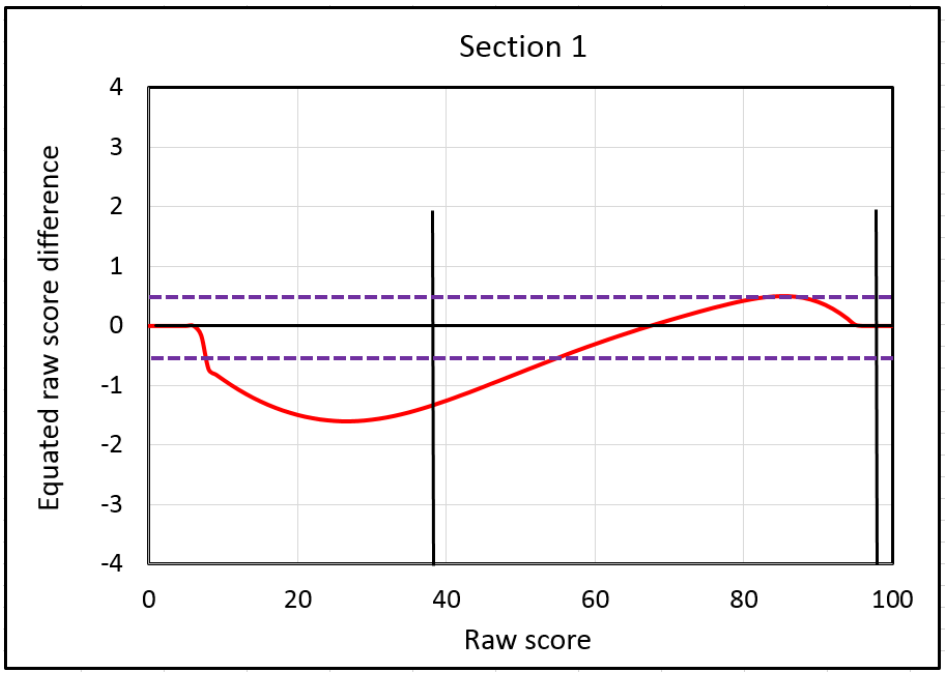

Figure 4. Equated raw score differences between Subgroup B conversion and Subgroup A conversion in Section 1.

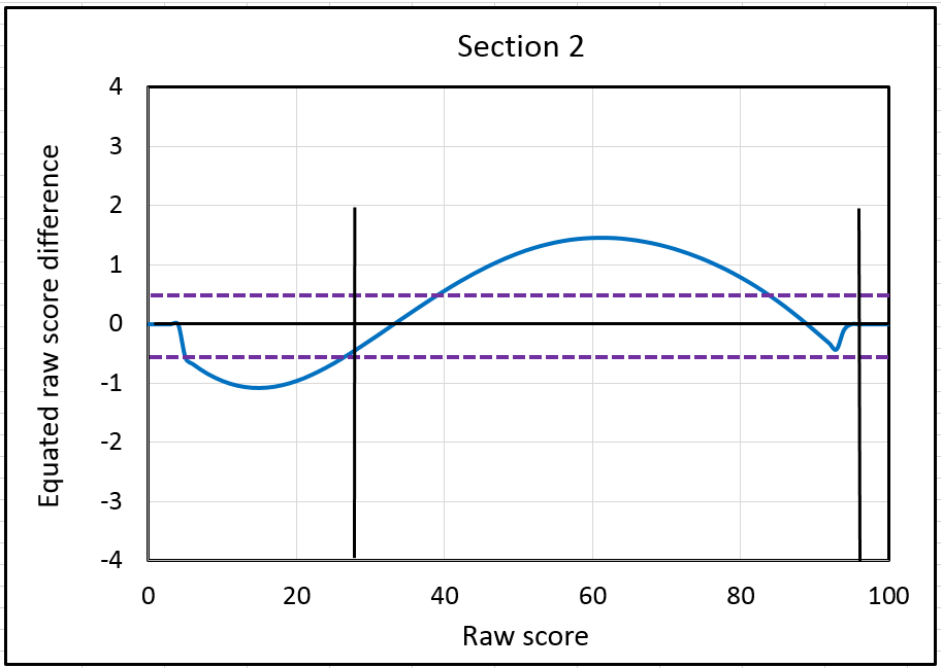

Figure 5. Equated raw score differences between Subgroup B conversion and Subgroup A conversion in Section 2.

Table 6 presents the means and standard deviations of the equated raw scores computed from Subgroup B test takers using each of the Subgroup A and Subgroup B conversions. On average, the raw score means slightly changed in both sections depending upon the subgroup conversions. When the Subgroup B conversion was applied to Subgroup B, the mean score was slightly higher than the one derived from the Subgroup A conversion. Although this trend was true in both sections, the magnitude of the mean difference was larger in Section 2 than in Section 1.

Table 6. Means and Standard Deviations of the Equated Raw Scores When the Conversions of Subgroup A and Subgroup B Were Applied to the Test Takers of Subgroup B

Section	Subgroup A *M*(*SD*)	Subgroup B *M*(*SD*)
1	79.47 (12.42)	79.65 (12.72)
2	69.70 (15.49)	70.62 (15.43)

Note. $N = 63{,}535$.

Table 7 presents a distribution of the equated raw score differences (scores derived from the Subgroup B conversion minus scores derived from the Subgroup A conversion) associated with each test taker of Subgroup B. In Section 1, the equated raw scores remained unchanged for approximately 90% of test takers. The differences were primarily within the range of −1 to +1 after rounding. In Section 2, however, the proportion of test takers whose scores remained the same was much smaller. More than 75% of test takers' scores increased by 1 point when the Subgroup B conversion was applied. This direction is consistent with the fact that many B and C DIF items were more difficult for Subgroup B than for Subgroup A. Even so, the differences were within the range of −1 to +1. None of the score differences was greater than 2 points.

Table 7. Distribution of Differences Between the Equated Raw Scores Derived From the Subgroup B Conversion and Those Derived From the Subgroup A Conversion When Applied to Subgroup B Test Takers

Difference	Section 1 (%)	Section 2 (%)
−2	0.42	0.0
−1	6.26	1.19
0	89.71	22.00
1	3.60	76.81
2	0.0	0.0

Note. The first column denotes Subgroup B conversion minus Subgroup A conversion. None of the difference scores was beyond the range −2 to +2. Calibration for both sections was separate. $N = 63{,}535$.

Discussion

When a test is administered to groups beyond the originally intended target group, testing programs desire to make similar inferences based on test scores. Even if the testing program conducted extensive research to support the test's purpose in the original population, more research would be needed to justify claims associated with the test's outcomes in its new context. The test used in this study offers a scenario similar to what might be encountered in practice: The test forms are built and equated based on data from Subgroup A. Test takers from Subgroup B take the same form, which is subject to the same scoring procedures as for Subgroup A. Score users interpret scores in the same way across the two subgroups, so that the accuracy of those interpretations rests on the invariance of the test's psychometric properties across the two subgroups.

We examined item-level invariance of the test using a DIF procedure. We looked for individual items that performed differently in the two subgroups. We also looked for any imbalance in the number and direction of DIF items. For a test equally suited to both subgroups, we would expect to see no DIF items (or at least balanced DIF across the two subgroups). For Sections 1 and 2, we did find differences in the performance of some items across the two subgroups, however. As the DIF analyses indicate, the apparent difficulty of several items decreased or increased depending upon subgroup membership in both sections, even after controlling for overall test score. In Section 1, however, items showing negative DIF (i.e., C− and B− DIF) cancelled the effects of items showing positive DIF (i.e., C+ and B+ DIF) without showing any systematic direction in bias at the test score level.

We evaluated test-level invariance using a test for subpopulation invariance of linking functions. We looked for differences in the score conversions based on the two subgroups. We also evaluated the percentage of people in a subgroup whose scores would change, and by how much, if the score conversion based on that particular subgroup were used. If test score invariance held, we would not expect individual test scores in Subgroup B to change if we were to use the score conversion based on Subgroup B as opposed to the normally used conversion based on Subgroup A. When Subgroup B's Section 1 conversion was applied to Subgroup B's test takers, most test takers' equated raw scores remained the same after rounding. Subpopulation linking invariance generally held across the score region where most Subgroup B test takers were located.

The same trends did not emerge for Section 2. Although Section 2 exhibited fewer C DIF items than did Section 1, about 77% of Subgroup B test takers' scores increased by 1 point, after rounding, when the Subgroup B conversion was applied to them. This indicates some inconsistency in the raw-to-equated raw score relationship as a function of subgroup membership. Score equity of Section 2 was not strictly satisfactory at the test score level. In reality, however, subpopulation linking invariance is never achieved absolutely. Instead, the question is whether subpopulation invariance holds closely enough such that the differences among subpopulation conversions are negligible and thus the differences will not adversely affect test takers (Dorans & Holland, 2000). In Section 2, although the degree of violation of score equity (maximum difference close to 1.5 raw score points) was larger than the DTM criterion (half a raw score point), the decision on whether this level of violation is large enough that test takers or score users would care must depend on the characteristics of the test (e.g., certification test or admission test).[4]

Lack of subpopulation invariance in a linking function indicates that some fundamental aspect of the test is not consistent across those subpopulations. When a test initially designed for a particular subpopulation is administered beyond this original target subpopulation, and when subpopulation invariance fails to hold, then we must question the equivalence of test scores across the two subpopulations. In other words, the meaning of test scores differs across subpopulations, and comparing the scores across the two subpopulations could be unfair.

When subpopulations are about equal in size, small differences in test performance across subpopulations might be mitigated by equating the test using the combined population. Kim and Robin (2017) emphasized that not using all test takers for item calibration and linking might still be defensible when within-group performance differences are large and when varying subpopulation proportions across test administrations greatly influence overall administration performance. According to the authors, using all examinees for statistical procedures required for reporting scores may add unwanted bias to the item bank due to the instability of linking samples; and more importantly, not all subpopulations are equally vulnerable to test security, which can be a threat to the validity of test scores. Inclusion of unstable subpopulations in the linking sample may exacerbate the situation by adding different sources of bias to test scores. Although their investigation indicated that failure to include two subpopulations in the operational calibration may have biased the resulting linking, they concluded that the

continuation of the current practice (using a major subpopulation of the test) seemed the safer choice for maintaining the stability of the reporting scale over time. Even so, it is generally assumed that any potential bias associated with the use of a particular subpopulation conversion would increase if the proportion of the excluded subpopulation(s) became larger over time.

When the exclusion of (nonmajor) subpopulations raises fairness concerns, two options can be considered for conducting psychometric procedures. One option is to apply proper weights derived from the testing population group to each administration to make the proportion of each subgroup similar across administrations. Using a certain weighting scheme based on the target population will help maintain sample stability over administrations. Another option is to implement a different calibration approach, such as *concurrent calibration with multigroups*. As the name indicates, this approach involves estimating item parameters using all data from all subgroups simultaneously to obtain a common IRT scale. Under concurrent calibration, Category A items (no DIF) can be treated as common items across subgroups, but the remaining items (i.e., Categories B and C) can be treated as unique items. Because the majority of items, which are Category A, take into account not only a major subpopulation's responses but also other (nonmajor) subpopulations' responses, the resulting conversion associated with the major subpopulation would be more likely to be appropriate for the entire population of test takers.

In this study, many C DIF items appeared in both sections. A testing program might choose to score those items (rather than to remove them) as long as the items were not flawed and there was no clear indication to explain the reason of difference. However, a more important question is how to use or revise those C DIF items in the future. Excluding problematic items from the item bank could be considered an option to reduce statistical bias in the resulting linking function. On the other hand, the presence of C DIF items does not necessarily signal bias. Rather, it indicates differential dimensionality of those items, which could be the result of item content. If a large proportion of items function differently across subgroups, the removal of those items from the item bank threatens the validity of the test due to the reduction in 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. If the items are retained in the test, it is suggested that they not be used as common items for linking and equating.

Additional evidence, such as the relationship between item type and item DIF, could be used in the process of not only item pool expansion but also test form assembly. A better understanding about the distinction among subgroups may enhance the development of an assessment, including carefully developed content specifications for the test and the writing review, tryout, and revision of test questions. It is important to plan for how individual items work together as a test. The item-level results based on data collected from a particular subpopulation of test takers may not be representative of the entire population of test takers.

Item and test invariance are necessary conditions for the fair use of test scores across multiple subpopulations. They should by no means be considered sufficient, however. Although this study illustrates a practical way to produce useful empirical evidence to evaluate the extended use of a test, it has several limitations, as follows. First, validation of test score uses is an ongoing process involving the score's relationship to external variables based on a sound theory about the proposed test use. Because predictive and concurrent validity claims were not the focus of the present study, we did not cover them, despite their importance. Second, we did not investigate the actual impact of item-level dependencies on score-level invariance. Following the current practice, we scored all items, including C DIF items, and used them in the process of IRT true score equating. It is expected that the direction of C DIF items might be crucial to determining score-level invariance. Lee and Zhang (2017) conducted an extensive simulation study to examine the effect of DIF on measurement consequences (e.g., total scores, ability estimates, test reliability, and standard error of ability estimation). In their simulation, the greatest DIF effect was less than 2 points on a 0–60 total score scale and about 0.15 on the IRT ability scale. A future study for assessing the impact of item-level invariance on test-level invariance in a systematic matter would be informative. Third, because we used a hypothetical reference form to compare the Subgroup B conversion to Subgroup A, it is worthwhile to note that the equating invariance findings observed in this study do not reflect actual empirical findings for the program from which the data were taken. Fourth, we examined a single form of the test for illustrative purposes. In practice, replication using multiple forms would be necessary to lead to solid conclusions about the extended use of the test.

The linking procedures used in this study might seem unconventional. Although classical DIF analyses have been used to evaluate common items in observed-score equating models (see Michaelides, 2008; von Davier, 2013), use of this methodology to evaluate common items in

TBLT is unusual. Still, the DIF analysis presents a practical solution to the current problem. This study illuminates how subpopulation linking invariance can be examined in such a way as to produce useful evidence to evaluate the appropriateness of expanded test use for new populations.

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.

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

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

Castellano, K. E., & Kolen, M. J. (2016). Comparing test scores across grade levels. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 135–155). New York, NY: Routledge.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum No. RM-94-10, pp. 93–124). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Educational Testing Service. (1988, November). *Procedures for use of differential item difficulty statistics in test development* [ETS internal memorandum]. Princeton, NJ: Author.

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing, 7*, 39–52. https://doi.org/10.1080/15305050709336857

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164–187. https://doi.org/10.1111/j.1745-3984.2001.tb01121.x

Hanson, B. A., & Zeng, L. (1995). ST: A computer program for IRT scale transformation [Computer software]. Retrieved from http://www.uiowa.edu/~casma/computer_programs.htm

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

Kim, S., & Robin, F. (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

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practice*. New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-4310-4

Lee, Y.-H., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing, 17*, 23–54. https://doi.org/10.1080/15305058.2016.1224888

Michaelides, M. P. (2008). An illustration of Mantel–Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research, and Evaluation, 13*(7), 1–16.

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. https://doi.org/10.1177/014662168300700208

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*, 605–623. https://doi.org/10.1007/s11336-013-9319-3

Wendler, C., & Powers, D. (2009, April). *What does it mean to repurpose a test?* (R&D Connections, No. 9). Princeton, NJ: Educational Testing Service.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Notes

[1] To make the hypothetical reference forms realistic, we based them on several existing forms. In Section 1, the mean of 100 items' *a* parameters was 0.60 (*SD* = 0.25; range, 0.09–1.19), and the mean of their *b* parameters was −0.98 (*SD* = 1.26; range, −5.74–1.74). In Section 2, the mean of 100 items' *a* parameters was 0.59 (*SD* = 0.25; range, 0.16–1.32), and the mean of their *b* parameters was −0.42 (*SD* = 1.03; range, −2.54–2.93).

[2] To refine a criterion score for the third round, a few newly designated C DIF items at the second round were also removed from the second-round criterion score. For that reason, 80 and 91 items were used to determine the criterion score in Section 1 and Section 2, respectively.

[3] Note that these results are hypothetical and for illustrative purposes only.

[4] The results reported here are hypothetical, so no one can be adversely affected. In an actual testing program, we would want to consider the uses to which the test is put and how large a difference could have meaningful consequences.