Differential Performance on National Exams: Evaluating Item and Bundle Functioning Methods using English, Mathematics, and Science Assessments

Syed Latifi¹, Okan Bulut¹, Mark Gierl¹, Thomas Christie², and Shehzad Jeeva³

Abstract

The purpose of this study is to evaluate two methodological perspectives of test fairness using a national Secondary School Certificate (SSC) examinations. SSC is a suit of multi-subject national qualification tests at Grade 10 level in South Asian countries, such as Bangladesh, India, and Pakistan. Because it is a high-stakes test, the fairness of SSC tests is a major concern among public and educational policy planners. This study is a first attempt to investigate test fairness of the national SSC examination of Pakistan using two independent differential item functioning (DIF) and differential bundle functioning (DBF) procedures. The SSC was evaluated for possible gender bias using multiple-choice tests in three core subjects, namely, English, Mathematics, and Physics. The study was conducted in two phases using explanatory item response model (EIRM) and Simultaneous Item Bias Test (SIBTEST). In Phase 1, test items were studied for DIF, and items with severe DIF were flagged in each subject. In Phase 2, the item bundles were analyzed for DBF. Three items were detected with large DIF, one for each subject, and one item bundle was detected with a negligible DBF. Taken together, the results demonstrate that there is no major threat to the validity of the interpretation of examinees’ test scores on the SSC examination. The outcome from this study provided evidence for test fairness, which will enhance test development practices at the national examination authorities.

Keywords

explanatory item response modeling, differential item functioning, differential bundle functioning, SIBTEST, validity, test fairness, national examinations

Differential item functioning (DIF) occurs when examinees who have the same ability but belong to different groups have a different probability of answering a test item correctly, after being controlled for overall ability on the construct measured by the test. DIF could be attributed as item bias or item impact. Item bias is a systematic error or invalidity in how a test item measures a construct for the members of a particular group (Gierl, Rogers, & Klinger, 1999). When the test item unfairly favors one group of examinees over another, the item is considered biased. Alternatively, group disparity on item performance due to actual knowledge and experience of a group, on the construct of interest, is called item impact. For example, the inclusion of map as part of the item stem may not require prior knowledge of content on a map but it could make an item easier or difficult for certain examinees, based on their prior knowledge about the places on the map (Ercikan, 2002). Differential bundle functioning (DBF) is a concept built upon DIF, in which a subset of items or testlets within a test are organized to form a group of two or more items. These testlets or item bundles are then analyzed for differential performance among the groups, after controlling for their overall ability. Specific organizing principles need to be followed while creating a bundle of items, such as grouping the items based on their content areas.

Researchers from different geographical contexts have presented explanations as to why DIF or DBF occurs. For example, a study from England (Ong, Williams, & Lamprianou, 2011, 2013) explained the gender DIF and DBF as a function of mental capability of processing, storing, and
retrieving the item solution. A study from Turkey (Kalaycioglu & Berberoglu, 2011) explained that the DIF in items can be due to item format characteristics, subject matter–related factors, and cognitive skills measured on the test. A study from China (Wei et al., 2012) has attributed gender DIF due to latent advantage of processing language and numbers. A study from Pakistan (Abida et al., 2011) has attributed DIF as a function of weak instructional and assessment practices.

Secondary school leaving examinations (e.g., Grade 10, high school diploma) are high-stakes assessments, and ensuring that these assessments are free from bias (DIF) is important from at least three ways. First, a fair test ensures that the examinees have attained a prescribed level of achievement. Second, a fair test enables the valid classification of examinees with reference to their ability on the test. Third, a fair test provides empirical evidence which could facilitate the critical evaluation of educational objectives, examination policies, and subject content, and, more broadly, could improve the methods of instructions. However, there exists a gap in literature about the possible differential performance in high-stakes South Asian context of Secondary School Certificate (SSC) examination. SSC is a suit of multi-subject national qualification tests at Grade 10 level in South Asian countries, such as Bangladesh, India, and Pakistan, which is equivalent to General Certificate of Secondary Education (GCSE) examination in England. In Pakistan, SSC is considered high stakes and, to qualify in SSC, each examinee has to pass at least eight subjects. If an examinee fails, then 2 years of academic training will be lost and the chances to continue education at the college and university level get reduced (Aly, 2007; Tinto, 1975). Therefore, DIF is an important topic for educators working in different geographical and educational contexts (Abida et al., 2011; French & Finch, 2015; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl et al., 1999; Rogers, Lin, & Rinaldi, 2011).

Measurement Models for DIF and DBF

Different measurement models could be used for studying DIF and DBF. These methods could use item response theory (IRT) or classical test theory (CTT). Methods based on IRT use the property of parameter invariance for identifying the DIF and DBF between the group of interests (e.g., gender, social background, native language, etc.), whereas methods based on CTT use the nonparametric techniques for identifying DIF and DBF between the group of interests. Different IRT-based methods could be used to detect item and person parameter invariance, for example, hierarchical generalized linear modeling (Kamata, 2001), extended structural equation modeling (B. O. Muthén, Kao, & Burstein, 1991), and explanatory item response modeling (EIRM; De Boeck & Wilson, 2004; Wilson, De Boeck, & Carstensen, 2008). These methods evaluate the invariance in item and person parameters as well as the interactions between item and person parameters that form the basis for DIF and DBF within the IRT framework. The DIF/DBF methods could be classified based on the procedure they used for matching the groups and on the assumptions they made for the item response function (IRF). The groups could be matched using observed score or using the estimates of the latent variable. When the parametric assumptions are made for a functional form of IRF, the procedure is called parametric, and nonparametric otherwise. The Mantel–Haenszel procedure (Holland & Thayer, 1988) and standardization method (Dorans & Kulick, 1986) are examples of observed-score nonparametric procedures, whereas the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993) is a latent-variable nonparametric procedure. Although many DIF detection procedures are available, a relatively small number of these methods are preferred based on their practicality as well as on their theoretical and empirical strengths (Gierl, Gotzmann, & Boughton, 2004; Shepard, Camilli, & Williams, 1985).

Different DIF detection methods flag DIF differently, and their results could complement each other or may refute otherwise. Ideally, multiple methods should be employed for simultaneous detection of DIF at the item and item–bundle level. Both EIRM and SIBTEST are capable of detecting DIF at item and bundle level; however, their agreement of detecting DIF was never evaluated in a high-stakes assessment settings. Grade 10 school exiting examination, such as SSC, is a high-stakes assessment in almost all geographical contexts of the world, and such examination often attracts high public scrutiny and accountability. To date, no attempt has been made to study the fairness of SSC examination using two statistically different differential functioning methods. Hence, the purpose of this study was to contrast EIRM and SIBTEST procedures for differential item and bundle functioning using high-stakes secondary English, Mathematics, and science assessments. Specifically, we evaluated the national SSC English, Mathematics, and Physics exams for gender DIF and tested whether any of the testlets created using subdomains within each subject show gender DBF. The outcomes from this methodological study could reveal the detection concordance between the EIRM and SIBTEST methods, as well as it will disclose whether the three national SSC assessments are free from fairness objections.

Theoretical Background

DIF Analysis Framework

The DIF analysis framework employs the concepts of primary and secondary dimensions to explain why DIF occurs. Dimension refers to a substantive characteristic of an item that can affect the probability of obtaining the correct response. Each item in a test is intended to measure the main construct called the primary dimension. DIF items measure at least one secondary dimension in addition to the primary dimension.
dimension (Ackerman, 1992; Boughton, Gierl, & Khaliq, 2000; Roussos & Stout, 1996a; Shealy & Stout, 1993). A secondary dimension is auxiliary if it is intentionally assessed or nuisance if there is no intended reason for its existence. DIF caused by auxiliary dimensions is benign and reflects impact, whereas DIF caused by nuisance dimensions is adverse and reflects bias. DIF is typically based on the comparison of two groups: the reference group, which is usually the majority group, and the focal group, which is usually a minority group. The focal group is compared with the reference group to detect bias in the items and item bundles.

While examining the item–bundle interaction between focal and reference groups, uniform and nonuniform DIF/DBF could occur. The **uniform** DIF exists when the amount of DIF between focal and reference groups is always constant. That is, the probability of answering an item correctly is greater for one group uniformly across all ability levels. On the contrary, **nonuniform** DIF occurs when there is an interaction between ability level and group membership. That is, the probability of answering an item correctly is nonuniform for examinee groups across levels of ability. In IRT terminology, nonuniform DIF is indicated by the crossing between the item characteristic curves of focal and reference groups. SIBTEST is equally powerful for detecting uniform and nonuniform DIF/DBF, whereas EIRM only allows for testing uniform DIF/DBF (French & Finch, 2015; Kan & Bulut, 2014; Narayanon & Swaminathan, 1996).

For this study, we have used the explanatory item response models (EIRMs; De Boeck & Wilson, 2004) and SIBTEST (Shealy & Stout, 1993). Both procedures are preferred methods based on their theoretical and empirical strengths (De Boeck et al., 2011; Gierl et al., 1999; Lamprianou, 2013; Rogers et al., 2011). The EIRM is particularly advantageous in the detection of DIF/DBF because it allows the estimation of both fixed and random DIF/DBF models. The fixed DIF/DBF model assumes that the amount of DIF/DBF is fixed across persons, while the amount of DIF/DBF is allowed to vary across persons in the random DIF/DBF model. The random DIF/DBF model is also known as random-weights DIF/DBF model (De Boeck & Wilson, 2004; Kan & Bulut, 2014).

By comparison, SIBTEST is particularly advantageous in situations when examinees’ prior knowledge of content (impact) is present in the data (Klockars & Lee, 2008), and its statistics could be compared against the well-established criteria of DIF classification. Moreover, SIBTEST has been found capable for adaptations to the multilevel data structures (French & Finch, 2015), and its DIF statistics have been found robust, compared with other nonparametric DIF detection procedures, when sample sizes for reference and focal groups are small (Klockars & Lee, 2008; Roussos & Stout, 1996b). Nevertheless, both EIRM and SIBTEST could account for differences in ability between the focal and reference groups, have a well-established statistical foundation, they are robust to different sample sizes, and could be used to evaluate items and item–bundles (Briggs, 2008; De Boeck et al., 2011; French & Finch, 2015; Lee, Cohen, & Toro, 2009; Lei & Li, 2013; Roussos & Stout, 1996a). These are conditions that are common to most high-stakes assessment settings.

**Explanatory Item Response Modeling**

All item response models (IRMs) contain at least one parameter to describe the item and at least one parameter to describe the person (Hambleton, Swaminathan, & Rogers, 1991), which could then be used for the measurement of item and person properties. EIRM is an extension of IRM that employs the generalized linear mixed modeling framework to explain the item properties, person properties, and the interaction between item and person properties, thereby providing a broader measurement framework than IRM (Briggs, 2008; Wilson et al., 2008). The term “explanatory” refers to the content and contextual variables that could be used to group the item and/or the person based on common characteristics.

In the EIRM framework, dichotomously scored responses are denoted as \( Y_{pi} = 0 \) or 1, where \( p \) is an index of person and \( i \) is an index for items. The expected value of \( Y_{pi} \) is represented by \( \pi_{pi} \), which follows a binomial distribution. The logit function is used to put the probability values of \( \pi_{pi} \) into a continuous scale between \(-\infty \) and \(+\infty \). Mathematically, the logit function for the probability of responding an item correctly could be represented as \( \eta_{pi} = \ln \left(\pi_{pi} / (1 - \pi_{pi})\right) \).

Using the notation from De Boeck et al. (2011), the one-parameter logistic (1PL) model can be formulated as

\[
\eta_{pi} = \theta_p + \sum_{k=1}^{K} \beta_k X_{ik},
\]

where \( X_{i0} = 1 \) for all items; \( X_{ik} \) is a diagonal matrix with \( X_{ik} = 1 \), if \( i = k \) (\( k = 1, \ldots, K \)); index \( k \) has the same range as index \( i \), and 0 otherwise; and \( \theta_p \) is the ability parameter for person \( p \) with a multivariate normal distribution \( \theta_p \sim N(0, \sigma_\theta^2) \), and \( \beta_k \) refers to the item easiness as opposed to item difficulty in the traditional IRT models. De Boeck et al. (2011) also suggested writing the model in a simpler form as

\[
\eta_{pi} = \theta_p + \beta, \tag{2}
\]

to highlight the plus sign as item easiness instead of item difficulty. Also, it should be noted that although the \( \beta \) parameter represents “item easiness,” it in fact represents item difficulty. The higher the item difficulty parameter is, the easier the item becomes to respond correctly. Furthermore, this simple model can be extended by incorporating the group membership parameters \( Z \), the group-by-item interaction parameter \( \delta^{(p)} \), and \( W_{pi} \) to represent the product of item and group indicators \( (W_{pi} = X_i \times Z) \). The extended version of EIRM is written as

\[
\eta_{pi} = \theta_p + \beta + \zeta_{focal} Z + \delta^{(p)} W_{pi}, \tag{3}
\]
Here, $\xi_{\text{focal}}$ represents the main effect of focal group in comparison with the reference group, and $\delta^{(k)}$ is the DIF parameter. If an item is administered to two groups and they performed equally, then the equation would become $\eta_{pi} = 0_p + \beta$ because $Z_{p,i} = Z_{\text{reference}} = 0$. Conversely, if, for item, the group-by-item interaction is found to be significantly different from zero, then the item would be flagged as exhibiting DIF.

Testing DBF in EIRM

A similar statistical model could be used for testing DBF within the EIRM framework, which involves testing the interaction between the item and person properties. The term “bundle” indicates a set of items grouped together because they share a common content dimension, cognitive similarity, or a common item structure. However, in the EIRM framework, the DBF is a more parsimonious model for detecting bias using the common characteristics of items (De Boeck & Wilson, 2004). Equation 3 could be extended for DBF by incorporating the paired values of $(p, i)$, where the pair $(p, i)$ has a value of 1 on predictor $h$ if person $p$ belongs to the focal group and item $i$ belongs to the bundle, and a value of 0 otherwise. Both DIF and DBF could be combined in one model as follows:

$$\eta_{pi} = \theta_p + \beta_i + \xi_{\text{focal}}Z_{(p,i)\text{focal}} + \sum_{h=1}^{H} \delta^{(h)}W_{(p,i)h}. \tag{4}$$

Here, $\theta_p$, $\beta_i$, and $\xi_{\text{focal}}$ have the same representation as before. However, $Z_{(p,i)\text{focal}} = 1$ for the focal group and 0 for the reference group with $W_{(p,i)h}$ as the person-by-item predictor $h$, defined in such a way that $W_{(p,i)h} = 1$ if both $Z_{(p,i)\text{focal}} = 1$ and either $X_{(p,i)k} = 1$ (for item-specific DIF) or $X_{(p,i)h} = 1$ (for item bundle/subset DIF), and $W_{(p,i)h} = 0$ otherwise, and $\delta^{(h)}$ as the corresponding DIF parameter (De Boeck et al., 2011).

To identify items and bundles that function differentially for male and female examinees, the lme4 package (Bates et al., 2014) in R (R Core Team, 2015) was used, with males as the reference group and females as the focal group. A statistically significant value that is positive indicates DIF/DBF against the focal group whereas a negative value indicates DIF against the reference group. Thus, the sign can be used to determine which group is favored. To classify the SIBTEST effect size $\hat{\beta}_{\text{uni}}$, guidelines recommended by Dorans (1989) were used in this study. The absolute values of the $\hat{\beta}_{\text{uni}}$ statistic less than 0.050 indicate negligible DIF (Level A), between 0.050 and 0.099 indicate moderate DIF (Level B), and 0.100 and above indicate large DIF (Level C). Based on these guidelines, items are classified as “A” (negligible DIF), “B” (moderate DIF), or “C” (large DIF). Also, to determine statistical significance, an alpha level of .05 was used. DIF literature also suggested other interpretation guidelines. For example, Jodoin and Gierl (2001) suggested a slightly different guideline for classifying DIF using logistic regression. However, the guidelines suggested by Dorans are also widely used and they are appropriate for the current study (M. Gierl, personal communication, January 6, 2016). More recently, Dorans’s guidelines for interpreting the SIBTEST effect size were used by Puhan, Boughton, and Kim (2007) for analyzing the differential performance for test delivery mode (i.e., paper-and-pencil vs. computer-based testing) in a high-stakes assessment situation.

SIBTEST

SIBTEST provides a measure of effect size, Beta-uni ($\hat{\beta}_{\text{uni}}$), and a statistical test for each item or bundle. In this approach, the complete latent space is viewed as multidi-dimensional, $(\theta, \eta)$, where $\theta$ is the primary dimension and $\eta$ is the secondary dimension. The statistical hypothesis tested by SIBTEST is $H_0: \hat{\beta}_{\text{uni}} = 0$ versus $H_1: \hat{\beta}_{\text{uni}} \neq 0$. If $H_0$ is not rejected, then $\eta = 0$ and there is no DIF.

$\beta_{\text{uni}}$ can also be used to estimate the magnitude of DIF in terms of an effect size. To operationalize this approach, items on the test are divided into the suspect subtest and the matching or valid subtest. The suspect subtest contains the items or bundle of items that are believed to measure both primary and secondary dimensions, whereas the matching subtest contains the items or bundle of items that are believed to measure only the primary dimension. The matching subtest places the reference and focal group examinees into subgroups at each score level so their performances on items from the suspect subtest can be compared. To estimate $\hat{\beta}_{\text{uni}}$, the weighted mean difference between the reference and focal groups on the suspect subtest item or bundle across the $K$ subgroups is calculated by

$$\hat{\beta}_{\text{uni}} = \frac{\sum_{k=0}^{K} p_k d_k}{\sum_{k=0}^{K} p_k}. \tag{5}$$

$p_k$ is the proportion of focal group examinees in subgroup $k$, and $d_k$ is the difference in the adjusted means of reference and focal groups on the studied subtest items, or bundle, in each subgroup $k$ (see also Jiang & Stout, 1998). A statistically significant value of $\hat{\beta}_{\text{uni}}$ that is positive indicates DIF against the focal group whereas a negative value indicates DIF against the reference group. Thus, the sign can be used to determine which group is favored.

To classify the SIBTEST effect size $\hat{\beta}_{\text{uni}}$, guidelines recommended by Dorans (1989) were used in this study. The absolute values of the $\hat{\beta}_{\text{uni}}$ statistic less than 0.050 indicate negligible DIF (Level A), between 0.050 and 0.099 indicate moderate DIF (Level B), and 0.100 and above indicate large DIF (Level C). Based on these guidelines, items are classified as “A” (negligible DIF), “B” (moderate DIF), or “C” (large DIF). Also, to determine statistical significance, an alpha level of .05 was used. DIF literature also suggested other interpretation guidelines. For example, Jodoin and Gierl (2001) suggested a slightly different guideline for classifying DIF using logistic regression. However, the guidelines suggested by Dorans are also widely used and they are appropriate for the current study (M. Gierl, personal communication, January 6, 2016).
empirical analysis. For the purpose of the present study, the bundles were created using the test specifications. The test specifications and content-wise item details for each examination are presented in Appendix A. In some cases, DIF at the item level may not be statistically significant but can be easily detectable using the bundle approach (Douglas, Roussos, & Stout, 1996; Gierl et al., 2001; Nandakumar, 1993; as cited in Gierl, 2005) because the combined effect across the set of items can amplify the group difference. Hence, DBF analysis is effective at identifying groups of items that function together to generate a group difference.

### Method

#### Data Source

The sample of this study represents Grade 10 students from the affiliated schools of a national examination board in Pakistan. The present study used test data from 103 randomly selected affiliated schools of the examination board, from annual SSC examination in year 2011. The question-level data for three subjects were extracted from an electronic database system, and the students’ identifiable information (e.g., name, school name, and geographical details) was removed before the data were released to the authors.

The subjects of English, Mathematics, and Physics examinations were chosen, not only because of their relative importance of the results obtained from these tests in determining career paths of the examinees but also because these subjects are generally preferred by students (Iqbal, Shahzad, & Sohail, 2010). Each test item was developed by the content specialists using item development guidelines and was reviewed for content representation and sensitivity by at least three subject experts. Each subject has two paper components: multiple-choice question (MCQ, Paper I) and constructed-response question (CRQ, Paper II). For the purpose of this study, only the MCQ portion of the exam was studied. The MCQ test for English is composed of 25 items, the Mathematics test includes 30 items, and the Physics test contains 25 items. All items were dichotomously scored, and the sample size details appeared in Table 1. All tests are based on National Curriculum guidelines (Government of Pakistan Ministry of Education, 2006a, 2006b, 2006c), which describe the Competencies, Standards, and Benchmarks for SSC assessments. The test of English comprises of two subcontent areas, whereas the tests of Mathematics and Physics comprise of three subcontent areas. The details related to the content area in each subject are presented in Appendix A.

#### Psychometric Characteristics

To study the psychometric characteristics of the three test forms, the classical test score theory indices were computed using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). As shown in Table 1, the number of females outnumbered the males by approximately 100 across the three subjects (103, 118, and 98, respectively). Using the independent t test, the mean scores of male and female examinees differed significantly only for English ($p < .05$), whereas there was no mean difference in Mathematics and Physics. However, the weak Cohen’s $d$ statistic of 0.21 for English suggested comparability among males and females, and thus overall the performance of males and females was comparable across three subjects.

The reliability of Mathematics examination is the highest (.89 for both male and females) while the reliabilities of the English and Physics examinations were relatively lower but

### Table 1. Psychometric Characteristics for the SSC English, Mathematics, and Physics Examinations.

| Characteristics               | English Male | English Female | Mathematics Male | Mathematics Female | Physics Male | Physics Female |
|-------------------------------|--------------|----------------|------------------|--------------------|--------------|----------------|
| No. of examinees              | 982          | 1,085          | 819              | 937                | 818          | 916            |
| No. of items                  | 25           | 25             | 30               | 30                 | 25           | 25             |
| $M$                           | 14.14        | 15.03          | 19.21            | 19.29              | 15.41        | 15.49          |
| $SD$                          | 4.27         | 4.21           | 6.04             | 5.88               | 4.13         | 4.07           |
| Skewness                      | -0.08        | -0.24          | -0.30            | -0.31              | -0.41        | -0.39          |
| Kurtosis                      | -0.45        | -0.40          | -0.65            | -0.67              | -0.55        | -0.48          |
| $M$ item difficulty           | 0.57         | 0.60           | 0.64             | 0.64               | 0.62         | 0.62           |
| $SD$ item difficulty          | 0.20         | 0.20           | 0.15             | 0.16               | 0.16         | 0.16           |
| $M$ item discriminationa      | 0.28         | 0.28           | 0.38             | 0.37               | 0.26         | 0.26           |
| $SD$ item discrimination      | 0.09         | 0.10           | 0.09             | 0.09               | 0.15         | 0.14           |
| Group-wise consistencyb       | 0.79         | 0.79           | 0.89             | 0.89               | 0.81         | 0.81           |
| Test internal consistencyb    | 0.76         |                | 0.87             |                    | 0.71         |                |

Note. SSC = Secondary School Certificate.

aItem-to-total Pearson correlation (point-biserial).
bAlpha reliability coefficient.
comparable (.79 and .81, respectively). The higher value of Cronbach’s alpha for Mathematics reflects the higher discrimination for the Mathematics items. This is likely due to the test length and/or due to more structured nature of subject Mathematics, compared with English and Physics.

The results presented in Table 1 indicate that the test developers were successful in minimizing gender difference at the overall observed-score level, and there was no difference in mean performance between reference (male) and focal (female) groups for English, Mathematics, and Physics examinations. Similarly, the conformance on test difficulty and discrimination statistics between male and female examinee groups confirms the comparability of performance across three subjects. These findings are important because they suggested that the gender groups are equivalent and there is no prior mean differences among groups on the trait being tested, and the overall performance is comparable at the test score level.

Next, the examinees’ ability estimates were computed using the `ranef` function within the `lme4` package (De Boeck et al., 2011). The accuracy of ability estimation using `ranef` was found near perfect in other studies (e.g., Lamprianou, 2013). Figure 1 represents the density plot of ability estimates for male and female examinees in three subjects. The difference between the ability distributions of male and female examinees in Mathematics and Physics tests was negligible. Similarly, for English, the difference is fairly small, but the ability distribution of male examinees has two small peaks on the ability scale. Taken together, the comparison of ability distributions indicates that both male and female examinees tend to perform equally across three subjects.

### Testing the Dimensionality of Data

Confirmatory factor analysis (CFA) was conducted using Mplus 6 (L. K. Muthén & Muthén, 1998-2015) to investigate whether English, Mathematics, and Physics subtests of the SSC hold a unidimensional latent structure. A one-factor CFA model was fit to each of the three subtests of the SSC. Goodness-of-fit criteria, including comparative fit index (CFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA), were used to evaluate model-data fit of the one-factor CFA model. CFI and TLI are incremental fit indices that range between 0 and 1, with values closer to 1 indicating good fit. RMSEA is an absolute fit index that is independent of sample size and thus performs well as an indicator of practical fit. For CFA models with categorical data, Hu and Bentler (1999) suggested that RMSEA < .06, TLI > .90, and CFI > .90 indicate good fit. As shown in Table 2, the dimensionality analysis indicated that the three subjects had acceptable levels of model-data fit based on all three

| Subject    | n     | χ²     | df  | χ²/df | CFI   | TLI   | RMSEA |
|------------|-------|--------|-----|-------|-------|-------|-------|
| English    | 2,067 | 471.083| 275 | 1.713 | 0.972 | 0.970 | 0.019 |
| Mathematics| 1,756 | 809.221| 405 | 1.998 | 0.973 | 0.971 | 0.024 |
| Physics    | 1,734 | 402.527| 275 | 1.464 | 0.979 | 0.978 | 0.016 |

Note. CFA = confirmatory factor analysis; SSC = Secondary School Certificate; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation.
model-fit criteria, suggesting evidence for the unidimensional structure of three SSC subjects. The results of the chi-square model-fit tests also support this finding. For each subtest, the scree test (Cattell, 1966) was also conducted (see Appendix B) and it also conforms with the findings of CFA.

Testing the Item and Model Fit

Next, the IRT item difficulty and chi-square item-fit statistics were computed using the BILOG-MG (du Toit, 2003; Zimowski et al., 2003). The RMSEA was also computed as a second measure of item fit as well as for easy interpretation of chi-square indices. The RMSEA was computed using the formula suggested by Tennant and Pallant (2012; see also Steiger, 1998). To interpret the RMSEA values, the guidelines suggested by MacCallum, Browne, and Sugawara (1996) were used, which suggested that the RMSEA value between 0.08 to 0.10 provides a mediocre fit and below 0.08 shows a good fit (Hooper, Coughlan, & Mullen, 2008).

Table 3 shows the item difficulties and fit indices for the items across English, Mathematics, and Physics test forms. The average difficulty of items on test forms of English, Mathematics, and Physics was −0.635, 0.755, and −0.847, respectively. The item-fit statistics suggested that, for each subject, almost all items had a good model fit. The only items that provided poor fit to the data were 14 and 21 in Physics that had RMSEA = 0.12.

After establishing the test reliability and comparability on the psychometric characteristics, the IRT ability estimates, the unidimensionality of data, and the model-data fit between male and female examinees’ groups on three tests, the next step was to analyze the differential item and bundle functioning.

EIRM Results

Under the EIRM framework, the item-level DIF analysis was initiated by estimating using the R implementation of generalized linear mixed-effects models (glimmer) within the lme4
The glmer function consists of a random component, a linear component, and a linking component. The examinees’ responses on items were considered as the random component, the test items and their interactions with the gender group form the linear component, and the Bernoulli/binomial distribution (logit) was used as a linking component. The glmer considers the first item functions as the reference item, and that all other item parameters are estimated as deviations from the first (De Boeck et al., 2011). Thus, to get the $\delta_{\beta_i}$ statistics for the first item, glmer was run twice. For each item, the sign of an estimate was used to determine the direction for DIF. Negative sign represents that the item favors females, and positive sign suggests that the item favors males.

SIBTEST Results

By comparison, the $\hat{\beta}_{\text{uni}}$ was estimated using the software program SIBTEST (Shealy & Stout, 1993). For this phase, each test item was considered as a suspect subtest whereas the remaining items were considered as the matching subtest. This process was then repeated for all the test items, one by one. The sign of $\hat{\beta}_{\text{uni}}$ determines the direction of DIF, with a negative $\hat{\beta}_{\text{uni}}$ value favoring females and a positive $\hat{\beta}_{\text{uni}}$ value favoring males. The DIF results are shown in Table 4.

Assessing the Level of DIF

To visualize, the DIF estimates from SIBTEST were also plotted. As shown in Figure 2, the region above/below the red dotted line represents the items with large DIF (Level C) in favor of male/female group, the items between the red and orange dotted lines represent the items with moderate DIF (Level B), and items between the orange dotted lines represent the negligible DIF (Level A).

First, the evaluation of Figure 2 suggested the even spread of values. That is, the items are evenly distributed among gender groups for all three subjects. Next, it also reveals that one

### Table 4. Comparison of DIF Results using EIRM and SIBTEST for English, Mathematics, and Physics Test Items.

| Item | English | | | Mathematics | | | Physics | | |
|------|---------|---|---|---------|---|---|---------|---|---|
|      | $\hat{\beta}$ | SE | Favors | Beta-uni | Favors | $\hat{\beta}$ | SE | Favors | Beta-uni | Favors | $\hat{\beta}$ | SE | Favors | Beta-uni | Favors |
| 1    | -0.501* | 0.164 | Female | -0.20 | 0.192 | Female | -0.24 | 0.061 | 0.157 | Female | 0.40 |
| 2    | -0.008 | 0.191 | Male | -0.12 | 0.224 | Male | -0.07 | 0.047 | 0.147 | Male | 0.68* |
| 3    | 0.043 | 0.172 | Female | -0.30 | 0.196 | Male | -0.056 | 0.0458 | 0.185 | Female | -0.031* |
| 4    | -0.082 | 0.161 | Female | -0.073* | 0.196 | Male | -0.14 | 0.0837 | 0.147 | Female | -0.150* |
| 5    | 0.318 | 0.166 | Male | 0.010 | 0.205 | Male | -0.018 | -0.259 | 0.147 | Female | -0.015 |
| 6    | -0.007 | 0.169 | Female | -0.37* | 0.195 | Female | 0.041 | 0.084 | 0.147 | Female | 0.029 |
| 7    | 0.381* | 0.164 | Male | 0.015 | 0.256 | Male | -0.005 | -0.384* | 0.160 | Female | -0.026 |
| 8    | 0.201 | 0.161 | Male | -0.006 | 0.135 | Male | -0.087* | -0.051 | 0.151 | Male | 0.057* |
| 9    | 0.222 | 0.166 | Male | 0.012 | 0.209 | Male | -0.013 | -0.56* | 0.155 | Female | -0.072* |
| 10   | 0.232 | 0.160 | Male | 0.004 | 0.022 | Male | -0.029 | -0.133 | 0.148 | Male | 0.013 |
| 11   | 0.09 | 0.167 | Female | -0.04* | 0.742* | Male | 0.105* | -0.239 | 0.148 | Male | 0.007 |
| 12   | 0.457* | 0.160 | Male | 0.053* | 0.295 | Male | -0.004 | -0.39* | 0.160 | Female | 0.020 |
| 13   | 0.583* | 0.160 | Male | 0.066* | 0.122 | Male | -0.018 | -0.159 | 0.151 | Male | 0.011 |
| 14   | 0.698* | 0.162 | Male | 0.100* | 0.435* | Male | 0.023 | -0.074 | 0.147 | Male | 0.026 |
| 15   | 0.061 | 0.160 | Male | -0.039 | 0.115 | Male | -0.008 | -0.065 | 0.177 | Male | 0.019 |
| 16   | 0.09 | 0.162 | Male | -0.030 | 0.061 | Male | -0.048* | -0.117 | 0.147 | Male | 0.025 |
| 17   | 1.36 | 0.160 | Male | -0.040 | 0.431* | Male | 0.029 | -0.292 | 0.156 | Male | 0.008 |
| 18   | 0.243 | 0.175 | Female | 0.002 | 0.183 | Male | -0.022 | -0.153 | 0.150 | Male | 0.022 |
| 19   | 0.296 | 0.161 | Female | 0.026 | -0.034 | Male | -0.036 | -0.157 | 0.171 | Male | 0.006 |
| 20   | 0.127 | 0.287 | Male | -0.001 | 0.247 | Male | 0.007 | -0.445* | 0.147 | Female | -0.061* |
| 21   | 0.042 | 0.160 | Female | -0.065* | 0.027 | Male | -0.045* | Female | -0.136 | 0.153 | Male | 0.016 |
| 22   | 0.231 | 0.160 | Male | 0.017 | 0.379 | Male | 0.035 | -0.358* | 0.150 | Female | -0.029 |
| 23   | -0.067 | 0.169 | Male | -0.029 | -0.046 | Male | -0.062* | Female | -0.094 | 0.155 | Male | 0.018 |
| 24   | 0.232 | 0.160 | Male | 0.015 | 0.417* | Male | 0.033 | -0.114 | 0.148 | Male | 0.018 |
| 25   | 0.501* | 0.161 | Male | 0.079* | 0.002 | 0.201 | -0.040* | Female | -0.061 | 0.157 | Male | 0.029 |
| 26   | 0.17 | 0.244 | Male | 0.016 | 0.115 | Male | 0.019 |
| 27   | 0.358 | 0.198 | Male | 0.019 |
| 29   | 0.53* | 0.194 | Male | 0.059* | 0.025 |
| 30   | 0.427* | 0.194 | Male | 0.040 |

Note. DIF = differential item functioning; EIRM = explanatory item response model; SIBTEST = Simultaneous Item Bias Test. *represents $p < .05$ as well as B and C level of DIF. †represents $p < .05$ only.
item in each subject had large DIF. Specifically, Item 14 in the English test and Item 11 in the Mathematics test favored males, whereas Item 4 in Physics test favored females.

As an alternate view, the items could be organized based on the content areas in each of the three subjects. As shown in Figure 3, there are two content areas in English—Listening Skills and Reading Skills—three subareas in Mathematics—Coordinate Geometry, Trigonometry, and Theorems; Fraction, Functions, and Algebraic Manipulation; and Linear and Quadratic Equations, Inequalities, and Graphs—and three subareas in Physics—Electronics, Telecom, and Radioactivity; Electrostatics, Current, and Magnetism; and Waves, Sound, and Optics. Figure 3 also suggested that the DIF statistics (β_{uni}) are evenly spread across subcontent areas, among each subject.

**DBF**

As Phase 2 of this study, the analysis for DBF was conducted. For each test, the test specification was used as an organization principle for forming the bundles of items (Gierl, 2005; Gierl et al., 2004). The test specification was used because it not only guides the assessment of dimensionality in a subject but also outlines the achievement domain associated with the content areas and cognitive skills (Gierl, 2005). The test specification and content-wise item details of this study are presented in Appendix A. Eight item bundles were created: two for English and three for Mathematics and Physics. Table 5 presents the indices from EIRM and SIBTEST DBF analyses. Among eight bundles across three subjects, EIRM did not detect any DBF. However, one bundle from English was found statistically significant (p < .05), in favor of females within the content area of Listening Skills, as flagged by the SIBTEST.

**Discussion and Conclusion**

Certificate, diploma, and other high-stakes examinations such as the SSC examination in South Asia will continue to be used for making important decisions about the examinees and thus will affect an individual’s career path. Hence, it is imperative that test items be free from any source of bias. The findings from this study suggested that, for the most part, the items and item–bundles from three core subjects did not display DIF and DBF. The psychometric characteristics of the tests were comparable for males and females across English, Mathematics, and Physics. First, the classical test indices were computed including internal consistency measure for evaluating the reliability of tests. Second, the score distribution was evaluated for comparability using ability estimated from 1PL model. Also, using the effect size measure (i.e., Cohen’s d statistic), the groups were found to be comparable on Mathematics and Physics. However, the weak effect size measure for English suggested that at the test level, the male and female examinees were essentially the same and that neither was favored. Before applying the IRT-based EIRM, the data were also checked and found unidimensional.
Both EIRM and SIBTEST detection procedures identified nearly the same number of items as DIF items; hence, a consistent pattern of DIF was displayed across both statistical procedures. However, if the guidelines for categorizing the DIF were also developed for EIRM, then this framework could be improved because it would include both a statistical test and an effect size measure for identifying DIF. EIRM identified one item in each subject area as possessing large (or Level C) DIF, with two of these items favoring males and one favoring females. Given the total number of items with less than Level C DIF items (e.g., 24 of the 25 English items did not have Level C DIF), no noticeable DIF was found, an indication that the test development practices remained successful in ensuring the test fairness because the majority of test items were found free from gender bias. Furthermore, the DBF analysis was conducted where the bundles of items were created using the Table of Specification for each examination. From the eight

**Table 5. Comparison of DBF using EIRM and SIBTEST, for English, Mathematics, and Physics Test Items.**

| Subject  | Content area                                      | No. of items | EIRM   | SIBTEST |
|----------|--------------------------------------------------|--------------|--------|---------|
| English  | Listening Skills                                 | 12           | 0.060  | −0.277* |
|          | Reading Skills                                   | 13           | −0.060 | 0.058   |
| Mathematics | Coordinate Geometry, Trigonometry, and Theorems | 16           | 0.101  | −0.071  |
|          | Fraction, Functions, and Algebraic Manipulation  | 4            | −0.101 | 0.080   |
|          | Linear and Quadratic Equations, Inequalities, and Graphs | 10       | 0.097  | −0.033  |
| Physics  | Electronics, Telecom, and Radioactivity          | 9            | −0.004 | 0.014   |
|          | Electrostatics, Current, and Magnetism           | 8            | 0.004  | 0.013   |
|          | Waves, Sound, and Optics                        | 8            | 0.012  | −0.002  |

*Significant at p < .05.

**Note.** DBF = differential bundle functioning; EIRM = explanatory item response model; SIBTEST = Simultaneous Item Bias Test.

---

**Figure 3.** Phase 2: Gender difference for items from SSC English, Mathematics, and Physics, May 2011 examinations, organized into bundles using the test specification.

*Note.** SSC = Secondary School Certificate; DIF = differential item functioning.
item bundles, only one was found statistically significant in favor of females.

The effect of gender in differential performance on content and cognitive skills was studied by Gierl, Bisanz, Bisanz, and Boughton (2003). They concluded that males perform better than females on items that require significant spatial processing and that females perform better than males on items requiring memorization (or memory recall). Furthermore, Kan and Bulut (2014) reported that the items presented as word problems were differentially easier for female examinees. As Listening Skills were tested using a recorded dialogue (which was played at the exam hall using CD/cassette players) and the examinees are expected to recall dialogue and answer the test items, the finding from the present study is consistent with other reported findings in the literature (Gierl et al., 2003; Kan & Bulut, 2014). In this case, the items may reflect impact, not bias. Moreover, a Level C item from Waves, Sounds, and Optic of Physics may be due to impact, as it is assessing the knowledge from ultrasound (see Appendix C), which female examinees are more likely to comprehend. This is also consistent with the earlier studies in which females were found to have higher attitude toward social implications of science than males (Anwer, Iqbal, & Harrison, 2012).

The Level B DIF items, which are by definition producing moderate DIF, were distributed evenly between male and female examinee groups. Moderate DIF is of little concern at this stage as it would be more challenging to explain than the large DIF. However, in large-scale testing programs, many DIF items are classified as moderate DIF (Gierl et al., 2004; Linn, 1993), whereas in this study, we only found 16% Level B items, with negligible composite effect (i.e., $\sum \beta_{b,c} \approx 0.01$). This finding also confirms that the tests were only assessing the primacy dimension and thus were fair overall.

The gender-specific difficulty estimates (male, female) for severe DIF items are also in concordance with the EIRM and SIBTEST outcomes. They were $(-1.138, -0.768)$ for Item 14 in English, $(-0.342, 0.112)$ for Item 11 in Mathematics, and $(0.328, -0.581)$ for Item 4 in Physics. Substantive review is needed to interpret the reason for the DIF for these three items. For example, for Physics, a plausible explanation may be that females outperformed males because females may have prior knowledge about the use of ultrasound, and thus females may be more proficient than males when answering Item 4; this DIF is due to systematic difference between actual performance of males and females and should be attributed as impact. The details about these Level C questions can be inspected in Appendix C. The Phase 1 DIF was repeated after dropping these three items, and this time no item was flagged as Level C DIF in any of the three subject tests.

Taken together, the results of this study indicate that there were only three items with Level C DIF in the studied subjects of English, Mathematics, and Physics examinations, and the item bundle for Listening Skills was found significantly favoring female. However, to date, no agreed guidelines exist to categorize the DIF and DBF estimate from EIRM, and DBF estimate from SIBTEST, and thus research is needed to identify and evaluate the guidelines for interpreting differential item and bundle functioning in EIRM framework. Furthermore, the results from the present study suggested that the small amount of DIF found does not confound the validity of the interpretation of the examinees’ test scores on three studied SSC subjects and, based on the statistical evidence, it is safe to assume that the test development practices produce items that are generally fair for both gender groups. However, the substantive review by the panel of content experts could further confirm these findings.

Finally, in this study, we have presented and compared two methods for assessing uniform and nonuniform DIF, by using the data from high-stakes examination settings. The methods presented in the study have a direct implication for practice in at least three ways. First, although the data from SSC context were used, the method presented in this study could be adopted in other high-stakes assessment situations. Second, assessing differential functioning at the level of item and bundle allows two layers of analyzing the test fairness, one at item level and other at subtest level. Third, the item–bundle approach could facilitate the discussion about the validity of construct because the bundle approach provides more cohesive unit of evidence compared with the differential performances at the item level. Thus, the explanations of DIF through DBF could facilitate test developer, test publishers, subject teachers, and generally to all those who directly or indirectly use scores from high-stakes assessments for making judgment about students, assessments, or both.

**Appendix A**

Test Specification for English, Mathematics, and Physics Test Forms.

| Subject | Subject content area | Item No. in test form | Total items |
|---------|----------------------|-----------------------|-------------|
| English | Listening Skills     | 1-12                  | 12          |
|         | Reading Skills       | 13-25                 | 13          |
| Mathematics | Coordinate  | 15-30                 | 16          |
|         | Geometry, Trigonometry, and Theorems | 2-5 | 4 |
|         | Fraction, Functions, and Algebraic Manipulation | 1,6-14 | 10 |
| Physics | Electronics, Telecom, and Radioactivity | 17-25 | 9 |
|         | Electrostatics, Current, and Magnetism | 9-16 | 8 |
|         | Waves, Sound, and Optics | 1-8 | 8 |
Appendix B
Scree plots of the SSC English (top), Mathematics (middle), and Physics (bottom) tests.

Appendix C
Test Items With Severe Bias.

| Subject     | Question                                                                 | Content area                        |
|-------------|---------------------------------------------------------------------------|-------------------------------------|
| English     | 14. According to the passage, the construction of the Pisa tower began in | Reading Skills                      |
|             | A. the middle of the Field of Miracles                                   |                                     |
|             | B. white marble                                                          |                                     |
|             | C. the 12th century                                                      |                                     |
|             | D. the period of Mussolini                                               |                                     |
| Mathematics | 11. The pair of points which lie on the straight line (AB) ↔ is (0,0) and | Linear and Quadratic Equations,     |
|             | A. (2,0)                                                                 | Inequalities, and Graphs             |
|             | B. (3,−2)                                                                |                                     |
|             | C. (−3,−3)                                                               |                                     |
|             | D. (−2,2)                                                                |                                     |
| Physics     | 4. Ultrasound is used for different purposes, which of the following is   | Waves, Sound, and Optics            |
|             | A. Detection of fault in engine                                           |                                     |
|             | B. Measure the depth of an ocean                                          |                                     |
|             | C. Diagnosis of different diseases                                        |                                     |
|             | D. Ranging and detection of airplanes                                     |                                     |

Acknowledgments
The authors would like to thank the Aga Khan University - Examination Board (AKUEB) for its support of this research. However, the authors are wholly responsible for the methods, procedures and interpretations expressed in this study. Our views do not necessarily reflect those of the AKUEB.

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 and/or authorship of this article.

References
Abida, K., Majoka, M. I., Munira, A., Azeem, M., Hussain, A., & Habib, Z. (2011). Understanding test item quality: The development of an unbiased test. *International Journal of Science in Society*, 2, 195-211.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91.
Aly, J. H. (2007). *Education in Pakistan: A white paper revised. Document to debate and finalize the National Education Policy*. Retrieved from http://planipolis.iiep.unesco.org/upload/Pakistan/Pakistan%20National%20Education%20Policy%20Review%20WhitePaper.pdf.

Anwer, M., Iqbal, H. M., & Harrison, C. (2012). Students’ attitude towards science: A case of Pakistan. *Pakistan Journal of Social & Clinical Psychology, 9*(2), 3-9.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singman, H., & Dai, B. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7). Retrieved from http://CRAN.R-project.org/package=lme4.

Boughton, K., Gierl, M. J., & Khaliq, S. (2000, May). *Differential bundle functioning on Mathematics and Science achievement tests: A small step toward understanding differential performance*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton, Alberta, Canada.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*, 89-118.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28.

De Boeck, P., & Wilson, M. (2004). *A framework for item response models*. New York, NY: Springer.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217-233.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*, 199-215.

French, B. F., & Finch, W. H. (2015). Transforming SIBTEST to account for multilevel data structures. *Journal of Educational Measurement, 52*, 159-180.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3-14.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in Mathematics: A demonstration of the DIF analysis paradigm. *Journal of Educational Measurement, 40*, 281-306.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.

Gierl, M. J., Gotzmann, A., & Boughton, K. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research, 45*, 353-376.

Government of Pakistan Ministry of Education. (2006a). *National curriculum for English language*. Retrieved from http://www.ibe.unesco.org/curricula/pakistan/pk_al_eng_2006_eng.pdf.

Government of Pakistan Ministry of Education. (2006b). *National curriculum for Mathematics*. Retrieved from http://www.ibe.unesco.org/curricula/pakistan/pk_al_mt_2006_eng.pdf.

Government of Pakistan Ministry of Education. (2006c). *National curriculum for Physics*. Retrieved from http://www.ibe.unesco.org/curricula/pakistan/pk_sc_psc_2006_eng.pdf.

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

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.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*, 53-60.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Iqbal, M. H., Shahzad, S., & Sohail, S. (2010). Gender differences in Pakistani high school students’ views about science. *Procedia: Social and Behavioral Sciences, 2*, 4689-4694.

Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kalaycioglu, D. B., & Berberoglu, G. (2011). Differential item functioning analysis of the Science and Mathematics items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment, 29*, 467-478.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in Mathematics assessments. *International Journal of Testing, 14*, 245-264.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*, 271-285.

Lamprianou, I. (2013). Application of single-level and multi-level Rasch models using the lme4 package. *Journal of Applied Measurement, 14*, 79-90.

Lee, Y. S., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review, 10*, 365-375.
Lei, P. W., & Li, H. (2013). Small-sample DIF estimation using SIBTEST. Cochran’s Z, and log-linear smoothing. *Applied Psychological Measurement, 37*, 397-416. doi:10.1177/0146621613478150

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-366). Hillsdale, NJ: Lawrence Erlbaum.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Muthén, B. O., Kao, C. F., & Burstyn, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to Mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user’s guide* (7th ed.). Los Angeles, CA: Author.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout’s test for DIF. *Journal of Educational Measurement, 16*, 159-176

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Ong, Y. M., Williams, J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in Mathematics assessment using differential bundle functioning. *International Journal of Testing, 11*, 271-293.

Ong, Y. M., Williams, J. S., & Lamprianou, I. (2013). Exploring differential bundle functioning in Mathematics by gender: The effect of hierarchical modelling. *International Journal of Research & Method in Education, 36*, 82-100.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment, 6*(3). Available from http://www.jtla.org

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rogers, W. T., Lin, J., & Rinaldi, C. (2011). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education, 24*, 39-70.

Roussos, L. A., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

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

Shepard, L. A., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 411-419.

Tennant, A., & Pallant, J. F. (2012). The root mean square error of approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions, 4*, 1348-1349.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*, 89-125.

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., & Zhou, X. (2012). Gender differences in children’s arithmetic performance are accounted for by gender differences in language abilities. *Psychological Science, 23*, 320-330.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 91-120). Göttingen, Germany: Hogrefe & Huber.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0) [Computer software]. Morrisville, IN: Scientific Software.

**Author Biographies**

**Syed Latifi** is research methodologist and doctoral candidate at the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. He has over 10 years of experience in high-stakes examinations, assessment engineering, program evaluation, and psychometrics. His current research interests include innovations in assessment, curriculum mapping, cognitive modeling, and Big data analytics.

**Okan Bulut** is an assistant professor in the Measurement, Evaluation, and Cognition program and a member of Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. Dr. Bulut’s current research interests include differential item functioning and measurement invariance in assessments, test reliability, Item Response Theory models, computerized adaptive testing, and technology-enhanced assessments.

**Mark Gierl** is professor of Educational Psychology and the director of the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. His specialization is educational and psychological testing, with an emphasis on the application of cognitive principles to assessment practices. His research is funded by the Social Sciences and Humanities Research Council of Canada. Dr. Gierl holds the Tier I Canada Research Chair in Educational Measurement.

**Thomas Christie** is senior consultant at the Adam Smith International, and a professor emeritus of the University of Manchester. He was the founding director (now retired) of Aga Khan University Examination Board. He has over 5 decades of experience in curriculum development, educational leadership, teacher training, program evaluation, and management. Dr. Christie has numerous peer-reviewed publications to his credit, and has published several reports on national educational policy and planning.

**Shehzad Jeeva** is director of Aga Khan University Examination Board. He is also a public speaker who promotes science education in Pakistan. Dr. Jeeva’s current research interests include science education, frameworks for middle school assessment, educational leadership, and innovations in testing.