

**Using a Multidimensionality-Based Framework to Identify and Interpret  
the Construct-Related Dimensions that Elicit Group Differences**

**Mark J. Gierl**

Centre for Research in Applied Measurement and Evaluation  
University of Alberta

DRAFT: February 11, 2004

Paper Presented at the Annual Meeting of the American  
Educational Research Association (AERA)

**San Diego, California, U.S.A.  
April 12-16, 2004**

## Abstract

The Roussos-Stout (1996) multidimensionality-based DIF analysis framework emphasizes a substantively-informed selection of items for both the matching and studied subtest based on the dimensions suspected of underlying the test data. By contrast, standard DIF practice encourages an exploratory search for matching subtest items based on purely statistical criteria, such as a failure to display DIF. Using two examples, we demonstrate that these two approaches lead to different interpretations about the occurrence of DIF in a test. It is argued that selecting a matching and studied subtest, as identified using the multidimensional framework, can lead to a more informed understanding of why DIF occurs.

### **Using a Multidimensionality-Based Framework to Identify and Interpret the Construct-Related Dimensions that Elicit Group Differences**

According to the authors of the *Standards for Educational and Psychological Testing* (1999), *bias* occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups (e.g., race, ethnicity, language, culture, gender, disability, or socio-economic status). Bias is often attributed to construct-irrelevant dimensions that differentially affect the test scores for different groups of examinees. Group differences can also be attributed to item *impact*. Impact occurs when construct-relevant dimensions differentially affect the tests scores for different groups of examinees. In this case, the item is a relevant measure of the target construct and the difference between the groups reflects a true difference on that construct. Differential item functioning (DIF) studies are designed to identify and interpret these construct-related dimensions using a combination of statistical and substantive analyses. The *statistical analysis* involves administering the test, matching members of the reference and focal group on a measure of ability derived from that test, and using statistical procedures to identify group differences on test items. An item exhibits DIF when examinees from the reference and focal groups differ, on average, in their probabilities of answering that item correctly, after controlling for ability. The *substantive analysis* builds on the statistical analysis because DIF items are often scrutinized by expert reviewers (e.g., test developers or content specialists) who attempt to identify the construct-related dimensions that produce group differences. A DIF item is considered biased when reviewers identify some dimension, deemed to be irrelevant to the construct measured by the test, that places one group of examinees at a disadvantage. Conversely, a DIF item displays impact when the dimension that differentiates the groups is judged to be relevant to the construct measured by the test.

Considerable progress has been made in the development and refinement of statistical methods for identifying items showing DIF (see reviews by Clauser & Mazor, 1998; Millsap, & Everson, 1993) but the development and refinement of substantive methods designed to aid with the interpretation of these items have lagged far behind (e.g., Bond, 1993; Camilli & Shepard, 1994, Englehard, Hansche, & Rutledge, 1990; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999, O'Neill & McPeck, 1993; Plake, 1980; Roussos & Stout, 1996; *Standards for Educational and Psychological Testing*, 1999; Stout, 2002; Sudweeks & Tolman,

1993). The traditional approach—subjecting items flagged with DIF analyses to the scrutiny of reviewers—has not been successful because the interpretations tend to be inconsistent with the DIF statistics or unreliable among reviewers. For example, Camilli and Shepard (1994) reported that, in their experience, as many as half of the items with “large” DIF in any one study might not be interpretable. Angoff (1993) noted: “It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values” (p. 19). Roussos and Stout (1996) reviewed the DIF literature and claimed, “attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure” (p. 360). The authors of the *Standards for Educational and Psychological Testing* (1999) concluded:

Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items. (p. 78)

This impasse represents a fundamental problem in the study of group differences using DIF methods.

Roussos and Stout (1996) proposed a *multidimensionality-based DIF analysis paradigm* to bridge the gap between statistical and substantive analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). The first stage is a substantive analysis where DIF hypotheses are generated. The second stage is a statistical analysis where the DIF hypotheses are tested. By combining statistical and substantive analyses in a multidimensional framework, researchers and practitioners can begin to systematically identify and interpret the construct-related dimensions that produce group differences using DIF methods.

The purpose of this paper is twofold: In the first section, we describe the Roussos and Stout (1996) DIF analysis framework. In the second section we illustrate, using two examples, how the DIF analysis framework can lead to more interpretable results about the dimensions that produce

group differences when compared with the traditional approach to DIF detection. We conclude with a summary and we highlight some implications for practice.

#### DIF Analysis Framework: An Overview

Roussos and Stout (1996) proposed a multidimensionality-based DIF analysis framework to link substantive and statistical analyses so researchers and practitioners can begin to systematically identify and study the sources of DIF. The DIF analysis framework is rooted in the Shealy and Stout (1993) multidimensional model for DIF (MMD), which serves as a theoretical basis for understanding how DIF occurs. A dimension is a substantive characteristic of an item that can affect the probability of a correct response. The main construct the test is intended to measure is the primary or target dimension. The MMD is based on two assumptions: (a) DIF items elicit at least one secondary dimension in addition to the primary dimension and (b) a difference exists between the two groups of interest in their conditional distributions on the secondary dimension, given a fixed value on the primary dimension. Roussos and Stout (1996) interpreted the secondary dimensions further. The secondary dimensions are *auxiliary* if they are intentionally assessed as part of the construct on the test, which implies the construct of interest contains multiple dimensions. DIF caused by auxiliary dimensions is *benign* (reflecting impact). Alternatively, the secondary dimensions are *nuisance* if they are unintentionally assessed as part of the construct on the test. DIF caused by nuisance dimensions is *adverse* (reflecting bias). On a test of mathematics achievement, for example, knowledge of mathematics might be a primary dimension, critical thinking might be an auxiliary secondary dimension, and testwiseness (i.e., using strategies to select the correct answer based on knowledge of test item characteristics) might be a nuisance secondary dimension. If a DIF item favors females and this difference can be attributed to the critical thinking auxiliary secondary dimension, when considered in isolation from the mathematics primary dimension, then DIF is considered benign. Alternatively, if a DIF item favors males and this difference can be attributed to the testwiseness nuisance dimension, when considered in isolation from the mathematics primary dimension, then DIF is considered adverse.

The Roussos-Stout DIF analysis (1996) framework is a two-stage procedure built on the foundation provided by the MMD. The first stage is a substantive analysis where the dimensional

structure of the test is evaluated and, based on this structure, where DIF hypotheses are generated. To decide whether the data contain distinct dimensions, *organizing principles* are used to identify single items or bundles of items that share certain characteristics. Four different organizing principles have been used to identify dimensions on tests (Ackerman et al., 2003; Douglas et al., 1996; Gierl et al., 2001; Roussos & Stout, 1996). First, *test specifications* can guide the assessment of dimensionality. Test specifications outline the achievement domain and help test developers obtain a representative sample of items from this domain. The specifications also guide item writing and help structure the final form of the test based on the content *and* cognitive domain that the test is designed to measure. Thus, a thorough analysis of the content areas measured by the test and the cognitive skills required by the examinees to solve the items may help identify a subsets of items that measure distinct dimensions associated with these content areas and cognitive skills (e.g., Gierl et al., 2001; Oshima, Raju, Flowers, & Slinde, 1996). Second, a *content analysis* can guide the assessment of dimensionality. For example, content specialists can review items and identify dimensions based on specific item content. A content analysis is guided by the professional experience of the reviewers. Two variations of content review can be used: specialists may use their experience and judgment to identify dimensions during an item review (e.g., Bolt et al., 1996; Douglas et al., 1996) or content-based judgments can be found in the literature to guide interpretation using well-known tests (e.g., Gierl & Bolt, 2003). Third, *psychological analyses* can guide dimensionality assessment when the hypothesized item structure is formulated from a psychological perspective. For example, a cognitive task analysis could be used to identify skills that characterize mathematics performance (e.g., Gallagher, De Lisi, Holst, McGillcuddy-De Lisi, Morely, & Cahalan, 2000). These cognitive skills could be identified and operationalized using test items to inform a dimensionality assessment (e.g., Gierl et al., in press). Fourth, *empirical analyses* can guide dimensionality assessment by using statistical methods to facilitate the identification of dimensions. Empirical approaches include, but are not limited to, factor analysis, cluster analysis, latent class analysis, and multidimensional scaling. The outcomes from these empirical approaches are then interpreted. This approach is substantive to the extent that the dimensions identified with the empirical procedures are, in fact, interpretable (cf., Douglas, Kim, Roussos, Stout, & Zhang,

1999; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995).

Once the dimensions are identified, they must be and interpreted as either the primary or secondary dimensions and, further, the secondary dimensions must be distinguished as auxiliary or nuisance. Then, the DIF hypotheses can be formulated to guide the study of group differences. The DIF hypotheses specify whether a single item or bundle of items designed to measure the primary dimension also measures a secondary dimension, thereby producing DIF across specific groups of examinees. DIF attributed to auxiliary secondary dimensions is benign whereas DIF attributed to nuisance secondary dimensions is adverse.

The second stage in the Roussos-Stout DIF analysis framework is statistically testing the dimensionality-based DIF hypotheses. The statistical analyses are used to see whether the organizing principles reveal distinct primary and secondary dimensions across the groups under study. SIBTEST is used to test DIF hypotheses and quantify the size of DIF (Stout & Roussos, 1995). To operationalize SIBTEST, items on the standardized test are divided into the matching and studied subtest based on the dimensions identified in the substantive analysis. The matching subtest contains items believed to measure only the primary dimension. This subtest should be an accurate measure of a unidimensional matching criterion because examinees in each subgroups are placed at the same score level so their performance on items from the studied subtest can be compared. Alternatively, the studied subtest contains items suspected of measuring the primary and secondary dimensions. In other words, the accuracy and interpretability of the statistical outcomes in second stage depend, in part, on the accuracy and interpretability of the substantive dimensionality analyses from the first stage. SIBTEST then uses differences in the expected scores conditional on primary dimension across groups to test for DIF. The method can be applied using either dichotomously-scored items (Shealy & Stout, 1993) or polytomously-scored items (Chang, Mazzeo, & Roussos, 1995). Since the approach under both item scoring conditions is basically the same, the more general case, as it applies to polytomous items, is described.

As a first step, SIBTEST estimates  $ES_R(\theta)$  and  $ES_F(\theta)$ , the expected score for a studied item conditional on the primary dimension  $\theta$  for the reference and focal group, respectively.

However, in place of  $\theta$ , SIBTEST uses total scores for a matching subtest of items. Then the expected item scores are estimated as

$$ES_R(t) = \sum_{k=1}^m kP_{Rk}(t) \quad \text{and} \quad ES_F(t) = \sum_{k=1}^m kP_{Fk}(t),$$

where  $P_{gk}(t)$  denotes the empirical proportion of examinees in group  $g$  that obtain score  $k$  on the studied item and have matching subtest score  $t$ .  $ES_R(t)$  and  $ES_F(t)$  contain bias due to measurement error in the matching subtest. The defining feature of SIBTEST is its use of a regression correction procedure that determines adjusted expected item scores  $ES_R^*(t)$  and  $ES_F^*(t)$ . These adjusted scores more accurately reflect examinees of equal ability levels across groups and, thus, are more meaningful for comparing group differences on the studied item. SIBTEST uses a weighted average difference of these adjusted scores (weighted by the proportion of examinees obtaining matching subtest score  $t$ ) to estimate a DIF index denoted as  $\widehat{\beta}_{UNI}$ , where

$$\widehat{\beta}_{UNI} = \sum_{t=1}^T ([ES_R^*(t) - ES_F^*(t)] \frac{N_R(t) - N_F(t)}{N}).$$

In this formula,  $T$  is the maximum score on the matching subtest,  $N_R(t)$  and  $N_F(t)$  are the numbers of examinees obtaining matching subtest score  $t$  from the reference and focal groups, respectively, and  $N$  is the total number of examinees. For large samples,  $\widehat{\beta}_{UNI}$  is approximately normal assuming a null hypothesis of no DIF, and the standard error of  $\widehat{\beta}_{UNI}$  is given as

$$\widehat{\sigma}_{\widehat{\beta}_{UNI}} = \left[ \sum_{t=0}^T \left( \frac{N_R(t) - N_F(t)}{N} \right)^2 \left( \frac{\widehat{\sigma}_{Rt}^2}{N_{Rt}} + \frac{\widehat{\sigma}_{Ft}^2}{N_{Ft}} \right) \right]^{1/2}.$$

The test statistic  $SIB = \frac{\widehat{\beta}_{UNI}}{\widehat{\sigma}_{\widehat{\beta}_{UNI}}}$  is evaluated against a standard normal distribution, and a null

hypothesis of no DIF is rejected whenever  $|SIB| > z_{1-\frac{\alpha}{2}}$ .

The multidimensional DIF framework has three noteworthy strengths. First, the framework is guided by a formal multidimensional model for understanding how DIF occurs. This framework emphasizes that a careful study of the underlying dimensions of a test is needed where a distinction is made between the primary, or intended *target* dimension of a test, and the secondary, or unintended auxiliary and/or nuisance dimensions of a test. Organizing principles guide the dimensionality assessment leading to a distinction between the primary and secondary dimensions. Then, DIF hypotheses are specified. Items or bundles of items that measure the secondary dimension should demonstrate a disproportionate difference between the reference and focal group relative to what should be observed on items that measure only the primary dimension, thereby producing DIF. Therefore, the studied subtest contains the item or bundle of items believed to measure the primary and secondary dimensions based on the substantive dimensionality-based analysis whereas the matching subtest contains items believed to measure only the primary dimension. Because of this important distinction between the primary and secondary dimensions, the multidimensional model can also guide the formation of the matching and studied subtests to improve DIF detection and interpretation.

Second, the framework draws on the confirmatory logic of hypothesis testing for DIF detection and interpretation. The confirmatory approach begins with substantive analyses designed to identify an interpretable dimensional structure and to generate DIF hypotheses about how groups may differ on this structure. It is followed with a statistical analysis where the dimensionality-based DIF hypotheses are tested. Organizing principles, which guide the dimensionality assessment, can also be used to identify and structure the matching and studied items. Then, SIBTEST is used to test the studied items. Each DIF study, therefore, provides a test of the proposed hypotheses. A confirmatory approach provides better Type I error control than a single-item exploratory DIF approach because only a comparatively small number of DIF hypotheses are tested. A confirmatory approach also has great potential to enable researchers and practitioners to systematically identify and interpret the sources of DIF so a body of *confirmed* DIF hypotheses can be created (Stout & Roussos, 1995). These confirmed hypotheses, accumulated over studies, may lead to a better understanding of why DIF occurs (Bolt, Froelich,

Habing, Hartz, Roussos, & Stout, 1999; Douglas et al., 1996; Gierl et al., 2001; Gierl & Khaliq, 2001; Gierl et al., in press; Stout, 2002).

Third, the framework can be used to evaluate single items *and* bundles of items. DIF hypotheses are specified and tested to determine whether items designed to measure the primary dimension also measure a secondary dimension, thereby producing DIF. In some cases, a single item may not yield an adequate measure of the secondary dimension that produces DIF (Douglas et al., 1996; Gierl et al., 2001; Nandakumar, 1993). But a bundle of items, by comparison, provides a broader sample of examinee performance over a larger sample of the secondary dimension and, therefore, may be easier to detect and interpret. Moreover, when these bundles tap a secondary dimension, they may *amplify* and detect group differences leading to a more powerful statistical analysis even when the same items tested separately show no statistically significant effects (Nandakumar, 1993).

#### Two Illustrative Examples: DIF Analysis Framework Compared to Traditional DIF Approach

To illustrate how the DIF analysis framework can promote the study of group differences, two examples are presented. In the first example, gender differences on a multidimensional reading comprehension passage are evaluated. In the second example, gender differences on a multidimensional test anxiety scale are assessed. In each example, the DIF analysis framework is compared with a more traditional approach to DIF detection to illustrate how these two approaches can lead to the selection of different matching and studied subtests and, consequently, different interpretations about group differences. We begin with a description of the traditional approach.

The traditional approach to DIF detection implies that each item is first tested statistically using a conditional DIF detection method and then scrutinized using some form of substantive review to identify the cause of the group difference (e.g., Camilli and Shepard, 1994; Ramsey, 1993; Zieky, 1993). This approach has also been described as an exploratory DIF analysis, meaning that items producing unexpected group differences are flagged statistically and then scrutinized by reviewers who attempt to understand why the item may be more difficult for one group of examinees (Roussos & Stout, 1996).

As noted earlier, accurate DIF detection requires the identification of a matching subtest to ensure that examinees are correctly compared. The DIF analysis framework draws on a two-stage approach to identify the matching and studied subtest using substantive and statistical analyses. The traditional approach, by comparison, often draws on a statistically-based iterative purification procedure to identify the matching subtest which, in turn, is used to identify DIF items on the studied subtest. Iterative purification is a procedure where DIF items are flagged, these flagged items are removed from the matching subtest using statistical criteria, and then the data are re-analyzed using the purified (i.e., DIF-free) matching subtest to flag additional DIF items (e.g., Allalouf, Hambleton, & Sireci, 1999; Camilli & Shepard, 1994; Dorans & Holland, 1993, pp. 60-61; Holland & Thayer, 1988; Lord, 1980). This procedure is designed to refine the matching subtest by automatically removing items flagged for DIF using purely statistical criteria because it is assumed that these items also measure a secondary dimension to the disadvantage of one group (Roussos & Stout, 1996, p. 357). However, little or no attempt is made to interpret the DIF items before they are removed from the matching subtest when iterative purification is used.

The matching subtest in the DIF analysis framework and the traditional approach are clearly related. Specifically, items that are "pure" measures of the primary dimension should display little or no DIF and thus would be identified as matching subtest items in both approaches. However, there are also some subtle, but important, differences between the two approaches. For instance, non-DIF items would be included in the matching subtest following iterative purification but should not automatically be regarded as candidates for the matching subtest in the DIF analysis approach. Indeed, if it is known or suspected that an item is not measuring the primary dimension by studying outcomes from previous DIF analyses, analyzing archival or existing test data, or formulating DIF hypotheses from test specifications, content reviews, or psychological studies, then this item may be excluded from the matching subtest, regardless of the magnitude of DIF. Also, DIF items are automatically excluded from the matching subtest when iterative purification is used because statistical outcomes supersede insights based on substantive reviews. However, some DIF items that are strongly believed to measure the primary dimension may be included in the matching subtest, regardless of the magnitude of DIF, when the DIF analysis framework is used. These differences, once again, highlight the importance of

conducting statistical and substantive analyses when deciding on which items should be included or excluded from the matching and studied subtests. These two approaches can also lead to different statistical and substantive outcomes, as illustrated in the next section.

*Gender Differences on a High School Reading Comprehension Construct*

In the first example, the responses for males and females on a reading comprehension construct were compared. The reading comprehension construct was evaluated on a high school exiting exam used in the Canadian province of Alberta. The exam mark contributes 50% to a student's final course grade. Questions on the exam are based on concepts, topics, and facts from the provincial curriculum. The exam contains a 70-item multiple-choice section and a separate, two-hour written-response section. Only the multiple-choice items were analyzed in the current study.

Using the DIF analysis framework, substantive and statistical analyses were conducted to identify and interpret the construct-related dimensions that could produce gender differences on the reading comprehension construct. For the current example, two reading comprehension passages were compared across gender using a content review as the organizing principle to guide the first stage substantive analysis. The first passage was based on an essay, published in 1989, in *Time* magazine by journalist Lance Morrow entitled, "Metaphors of the World, Unite!". Ten items were used to evaluate students' understanding of this passage. The second passage was based on a play by William Butler Yeats, first published in 1934, called "Cathleen ni Houlihan". Twelve items were used to evaluate students' understanding of this passage. The passages and test items are presented in the Appendix. The dimensional structure of the items associated with the two passages was evaluated using DIMTEST, with the refined bias correction method (Froelich, 2000; also see Froelich & Habing, 2001), for a random sample of 2000 students (1000 males and 1000 females). To operationalize DIMTEST, the items for each passage were divided into the assessment subtest (AT) and the partitioning subtest (PT). Using the play items as AT, the essay items on PT formed a dimensionality distinct item set,  $T = 4.19, p < 0.01$ , demonstrating that the items on each passage tapped a distinct dimension. Then, DIF hypothesis were specified. In this example, no systematic gender differences were expected on the reading comprehension dimensions because we had no a priori evidence to

suggest that the items associated with each passage would favor either males or females. Moreover, the two passages, while dimensionality distinct, were believed to measure a legitimate multidimensional reading comprehension construct because the passages only differed by their content. A summary of the two reading comprehension passages is presented in Table 1.

For the second stage statistical analysis, the response data from the random sample of 1000 Grade 12 males and 1000 females were compared. Using the outcomes from the first stage substantive analysis, items on the matching and studied subtest were categorized using the dimensions associated with the essay and the play. The play items served as the primary dimension and the essay items served as the *auxiliary* secondary dimension. In other words, both dimensions were thought to measure a legitimate and important component of the reading comprehension construct. In fact, the play dimension was selected as the matching subtest only because it contained a larger number of test items which could improve the accuracy of the matching process. The DIF hypothesis, as previously described, specified that no systematic gender differences would occur. Table 2 (under the heading *Confirmatory*) illustrates how the items from the play performed when studied as a stand-alone matching subset using SIBTEST. The matching subtest contained one DIF item. In addition to this appealing empirical evidence, these items are conceptually appealing because they measure and match examinees on a distinct dimension associated with the reading comprehension construct, as measured by the diploma exam. Table 2 also contains the DIF results for each essay item. In this case, each  $\hat{\beta}_{UNI}$  represents the expected score difference for males and females on the essay items matched according to their performance levels on the dimension measured by the play reading-comprehension passage. The substantively-driven, DIF analysis procedure identified three DIF items and indicated that, overall, DIF favored males for the essay dimension but the difference was not statistically, as the expected essay score difference was only 0.035 points higher for males conditional on the play reading-comprehension passage. In sum, the DIF analysis framework indicates that there are not systematic gender differences on the passage-based reading comprehension dimensions measured by the high school exiting exam.

The traditional approach, which uses iterative purification to identify the matching subtest, produced markedly different results. To purify the data, items were analyzed for DIF using a

single-item analysis and any items with large DIF (i.e.,  $|\widehat{\beta}_{UNI}| \geq 0.050$  and  $p < 0.05$ )<sup>1</sup> were removed. Then, the purified matching subtest was used for another DIF analysis. Results are presented under the heading *Exploratory* in Table 2. The matching subtest contained 18 items measuring both the dimensions associated with the essay (7 items) and play (1 item). Because limited amounts of DIF were found in these items, their use in the matching subtest would appear to be supported statistically. This matching subtest, in turn, was used to test four studied items from the essay (3 items) and the play (1 item). Taken together, the studied items implied an expected score difference of 0.132 points higher for females conditional on the matching subtest, which was a statistically significant difference.

What has been ignored using the traditional approach, however, is the validity of the items included on the matching subtest, as they reflect the dimensions associated with reading comprehension. As has been emphasized in this paper, such decisions should be based on substantive and statistical outcomes, especially considering that iterative purification is not guaranteed to produce a interpretable partitioning of items for the matching and studied subtest. Using the DIF analysis framework, items were placed in the matching and studied subtests based on the dimensional structure of the test. These dimensions were not expected to produce gender differences. Using the traditional approach, on the other hand, items were placed in the matching and studied subtest based solely on the outcomes from a statistically-based iterative purification procedure. As a result, the matching and studied subtest contained a mix of items from both reading-comprehension passages. These two approaches produced different results and led to different interpretations about the nature of gender differences on the two passages. For example, at the item level, one item favoring males in the DIF analysis approach (item 5) changed to favor females in the traditional approach. At the bundle level, a more dramatic difference occurred. The DIF analysis approach produced a small, nonsignificant expected score difference of 0.035 points favoring males on the essay subtest conditional on the play subtest, which was consistent with the hypothesis that these dimensions would not produce systematic

---

<sup>1</sup> These guidelines for interpreting single-item DIF by combining the SIB statistical results with values for the  $\widehat{\beta}_{UNI}$  parameter estimate were proposed by Nandakumar (1993) and Roussos (personal communication, October 28, 1999).

gender differences. The traditional approach, on the other hand, produced a large, significant expected score difference of 0.132 points favoring females on the statistically-derived studied subtest conditional on the matching subtest.

*Gender Differences on a Test Anxiety Construct*

In the second example, data from the Test Anxiety Inventory (TAI; Spielberger, Gonzalez, Taylor, Anton, Algaze, Ross, & Westberry, 1980) were used. The TAI is a 20-item scale designed to measure individual differences in test anxiety. According to the authors, people who are high in test anxiety perceive evaluation situations as threatening and, as a result, are usually tense, apprehensive, nervous, and aroused in these situations. To assess this response, students report how frequently they experience symptoms of anxiety in testing situations using a 4-point scale where 1 indicates “Almost Never” and 4 indicates “Almost Always”. The TAI yields an overall test anxiety score, which is the sum of the 20 items. It also yields two subscale scores which measure emotionality and worry, two components of test anxiety identified by Liebert and Morris (1967). Emotionality is defined as autonomous nervous system reactions evoked by tests. Worry is defined as concern for the anticipated consequences of test failures. Each subscale contains eight items (thus, four items on the TAI are associated with neither the emotionality nor the worry dimensions). The factor structure associated with the emotionality and worry items is well established, as it has been replicated by researchers across different samples and age groups (Benson & Tippets, 1990; Everson, Millsap, & Rodriguez, 1991; Gierl & Rogers, 1996; Spielberger et al., 1980; Ware, Galassi, & Dew, 1990).

Using the DIF analysis framework, substantive and statistical analyses were conducted to identify and interpret the construct-related dimensions that could produce gender differences on the test anxiety construct. For this example, the two well-documented dimensions that produce test anxiety were compared across gender. Moreover, previous research indicates that emotionality items elicit higher item-level scores for females relative to worry items resulting in higher emotionality subscale scores (e.g., Benson & Tippets, 1990; Everson et al., 1991; Gierl & Rogers, 1996; Spielberger et al., 1980). As a result, the DIF hypothesis specified that items from the emotionality dimension would systematically favor females. The TAI items are presented in Table 3.

For the second stage statistical analysis, the response data from the random sample of 335 Grade 12 males and 389 females were compared. Using the outcomes from the first stage substantive analysis, items on the matching and studied subtest were categorized using the dimensions associated with worry and emotionality. The worry items served as the primary dimension and the emotionality items served as the *auxiliary* secondary dimension. That is, both dimensions were thought to measure a legitimate and important component of the test anxiety construct. The DIF hypothesis specified that the emotionality dimension would systematic favor females. Table 4 (under the heading *Confirmatory*) illustrates how the worry items perform when studied as a stand-alone matching subtest using Poly-SIBTEST. The matching subtest contained four DIF items. Despite this outcome, these items remain conceptually appealing because they measure and match examinees on a well-understood dimension closely related to test anxiety. Table 4 also contains the DIF results for each emotionality item. For these items, each  $\widehat{\beta}_{UNI}$  represents the expected emotionality score difference for males and females matched according to their levels on the worry dimension of test anxiety. The substantively-driven, DIF analysis procedure identified four DIF items, and indicated that, overall, the emotionality dimension produced an expected TAI score that was 0.342 points higher for females conditional on the matching worry subtest score, which was a statistically significant difference. In sum, the DIF analysis framework indicates that are systematic gender differences favoring females on the emotionality dimension of the test anxiety inventory.

The traditional approach, which uses iterative purification to identify the matching subtest, produced a different result. To purify the data, items were analyzed for DIF using a single-item analysis and any items with large DIF (i.e.,  $|\widehat{\beta}_{UNI}| \geq 0.100$  and  $p < 0.05$ )<sup>2</sup> were removed. Then, the purified matching subtest was used for another DIF analysis. Results are presented under the heading *Exploratory* in Table 4. The matching subtest contained 14 items measuring both the worry (4 items) and emotionality (7 items) dimension. Despite this combination of items from both constructs, their use in the matching subtest would appear to be supported statistically because

---

<sup>2</sup> Different guidelines were used to identify DIF items in the Poly-SIBTEST analysis when compared with the SIBTEST analysis, reported in the previous example, because the TAI items are scored on a four category scale.

limited amounts of DIF were found. The matching subtest, in turn, was used to test six studied items measuring both the worry (4 items) and emotionality (1 item) dimensions. Taken together, the studied items implied an expected TAI score 0.345 points higher for males conditional on the matching subtest score, which was a statistically significant difference.

Again, however, the validity of the items included in the matching subtest, as they reflect the intended construct of text anxiety, has been ignored. Using the DIF analysis framework, items on the TAI were placed on the matching and studied subtests based on the dimensional structure of the test and with the expectation, based on previous research, that emotionality items would elicit larger differences for females. Using the traditional approach, items were placed in the matching and studied subtests based solely on the outcomes from a statistically-based iterative purification procedure. As a result, the subtests contained a mix of both emotionality and worry items. These two approaches resulted in different conclusions about the nature of gender differences on the TAI. At the item level, the results across the two DIF detection procedures were unstable: seven items that favored one group with the DIF analysis approach favored the other group using the traditional approach (items 2, 8, 9, 11, 17, 18, and 20). At the bundle level, the unstable item-level differences resulted in a striking difference. The DIF analysis approach produced a large, significant expected score difference of 0.342 points in favor of females on the emotionality subtest conditional on the worry subtest score, which is consistent with previous research on gender differences with the TAI. The traditional approach produced the opposite outcome: Males had an expected score difference of 0.345 points higher than females using the statistically-derived studied subtest conditional on the matching subtest.

### Summary

Differential item functioning studies are designed to identify and interpret construct-related dimensions that elicit group differences (*Standards for Educational and Psychological Testing*, 1999). Roussos and Stout (1996) proposed the multidimensionality-based DIF analysis paradigm to unify the substantive and statistical approach to DIF detection because many researchers and practitioners reported that the outcomes from DIF statistical analyses were not interpretable. The first stage in this framework is a substantive analysis where the dimensional structure of the test is evaluated and where DIF hypotheses are generated. To decide whether the data contain

distinct dimensions, organizing principles are used to identify items or bundles of items that share specific characteristics. The DIF hypothesis specifies whether an item or bundle designed to measure the primary dimension also measures a secondary dimension, thereby producing group differences. The second stage is statistically testing the dimensionally-based DIF hypotheses. Statistical analyses are used to see whether the data, so structured using the organizing principle, reveal distinct primary and secondary dimensions which, in turn, elicit group differences consistent with the DIF hypotheses. The DIF analysis framework is guided by a formal multidimensional model for understanding how and why DIF occurs; it relies on the confirmatory logic of hypothesis testing, which increases the interpretability of items that display DIF; and it can be used to identify and interpret single items and bundles of items. Moreover, by combining substantive and statistical analyses in a unified framework, researchers and practitioners can identify the dimensions that elicit DIF, thereby accumulating results over studies to increase our understanding of which dimensions consistently elicit group differences.

#### Implications for Practice

The outcomes from this study have at least two implications for researchers and practitioners who study group differences. First, dimensionality analyses should be conducted to identify the constructs that influence examinees' test performance. Currently, the practice of conducting dimensionality analyses as part of a DIF study is rare (Roussos & Stout, 1996). Yet the outcomes from dimensionality analyses have clear benefits: they provide some indication of the dimensional structure for the test and they forced researchers and practitioners to consider how these dimensions could affect examinee performance. Moreover, as Zhang and Stout (1999) recently noted, "a certain pattern of separated clusters of items about the test composite should typically result from the categorical nature of many test specifications" (p. 214). This statement suggests that many tests could measure a multidimensional composite by virtue of the test development process and, therefore, a thorough analysis of the test specifications should be conducted as part of the dimensionality assessment. It might also be useful to conduct this substantive analysis with the help of specialists who have extensive knowledge of the content areas measured by the tests as well as the knowledge and cognitive skills required by examinees to solve the test items. If subsets of items clearly measure different content areas and/or

cognitive skills, then these items could measure distinct dimensions and may elicit group differences.

But identifying and interpreting the dimensions measured by a test is only part of the substantive analysis in the DIF analysis framework. DIF hypotheses must also be formulated. The DIF hypotheses specify whether a difference exists between the two groups of interest in their conditional distributions on the secondary dimension, given a fixed value on the primary dimension for a single item or bundle of items. In other words, these hypotheses specify how the construct-related dimensions could produce group differences. Thus, the analyses outlined in the DIF analysis framework become one important step in the much larger process of construct validation. For example, the authors of the *Standards for Educational and Psychological Testing* (1999) claim that bias may occur when various construct-related dimensions differentially influence group performance (see Standards 7.1 and 7.2). DIF hypotheses specify whether these dimensions are, in fact, relevant to the intended construct producing benign DIF due to auxiliary secondary dimensions or irrelevant to the intended construct thereby producing *adverse DIF* or *bias* due to nuisance secondary dimensions. These hypotheses, derived from empirical evidence, existing literature, and/or logical analyses, guide the study of group differences and contribute to test validation process by encouraging researchers and practitioners to identify and interpret the dimensions that affect test performance.

Second, researchers and practitioners must attend to both the matching and the studies subtests. Often, the focus is on the studied subtest (i.e., DIF items) without considering carefully the matching subtest. The DIF analysis framework guides the development of the matching and studied subtest. The matching subtest serves as a unidimensional criterion designed to place examinees at the same score level so their performance on the studied subtest can be compared. Conversely, the studied subtest contains items suspected of measuring a multidimensional criterion based on the substantive analysis. A DIF analysis is then conducted to compare the studied subtest relative to matching subtest to determine whether the secondary dimension can be detected. If the secondary dimension is detected, it can be attributed to the interpretable characteristics identified in the substantive analysis. The traditional approach, by comparison, applies statistical and substantive methods to the evaluation of items on the studied subtest but

only statistical methods—typically, using iterative purification—to the evaluation of items on the matching subtest. Moreover, little or no attempt is made to interpret DIF before items are removed from the matching subtest. However, iterative purification is not guaranteed to produce an interpretable partitioning of items on either the matching or the studied subtest. In the two examples presented in this paper, iterative purification was used to identify studied items with large DIF statistical values relative to a matching subtest with items displaying small DIF statistical values. However, in both cases, iterative purification also produced subtests that were multidimensional (i.e., the subtests contained items from both dimensions on the test). As a result, the subtests did not serve as interpretable dimensions for either the reading comprehension or test anxiety constructs. These examples are intended to demonstrate that researchers and practitioners who want to identify and interpret DIF items should attend to both the matching and studied subtest because these subtests must be validated empirically as meaningful comparative criteria in the study of group differences: We contend that this form of validation requires comprehensive statistical *and* substantive analyses.

## References

- Ackerman, T. A., Gierl, M. J., & Walker C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*, 37-53.
- Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36*, 185-198.
- Angoff, W. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.
- Benson, J., & Tippetts, E. (1990). Confirmatory factor analysis of the Test Anxiety Inventory. In C. D. Spielberger & R. Diaz-Guerrero (Eds.), *Cross-cultural anxiety* (Vol. 4, 149-156). New York: Hemisphere/Taylor-Francis.
- Bolt, D., Froelich, A., Habing, B., Hartz, S., Roussos, L., & Stout, W. (1999, September). *An applied and foundational research project addressing DIF, impact, and equity with applications for ETS test development*. Princeton, NJ: Educational Testing Service.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park: Sage.
- Chang, H.H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.
- Douglas, J., Roussos, L., and Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465-484.
- Douglas, J. Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October, 1992, administrations*. (Statistical Report No. 95-05). Newton, PA: Law School Admission Council.

- Englehard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*, 347-360.
- Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51*, 243-251.
- Froelich, A. G. (2000). *Assessing the unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.
- Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-190.
- Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (in press). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement*.
- Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.
- Gierl, M. J., & Bolt, D. (2003, April). *Implications of the multidimensionality-based DIF analysis framework for selecting a matching and studied subtest*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164-187.
- Gierl, M. J., & Rogers, W. T. (1996). A confirmatory factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement, 56*, 315-324.

- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- Hamilton, L., Nussbaum, M., Kupermintz, H., Kerkhoven, J., & Snow, R. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 science achievement. *American Educational Research Journal, 32*, 555-581.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Kupermintz, H., Ennis, M., Hamilton, L., Talbert, J., & Snow, R. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 mathematics achievement. *American Educational Research Journal, 32*, 524-554.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16*, 159-176.
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.
- Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*, 353-369.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. *Educational and Psychological Measurement, 40*, 397-404.

- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.
- 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.
- Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Preliminary Profession Manual for the Test Anxiety Inventory*. Redwood City, CA: Mind Garden.
- Standards for Educational and Psychological Testing*. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485-518.
- Stout, W. & Roussos, L. (1995). *SIBTEST manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.
- Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30*, 3-19.
- Ware, W. B., Galassi, J. P., & Dew, K. M. H. (1990). The Test Anxiety Inventory: A confirmatory factor analysis. *Anxiety Research, 3*, 205-212.
- Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 231-249.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Author Notes

Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5; Email: [mark.gierl@ualberta.ca](mailto:mark.gierl@ualberta.ca)

Table 1.

*Items on the Two Reading Comprehension Passages*

Item	Dimension
1. Use of phrase “elegant game”	Time Magazine Essay
2. Author’s tone	Time Magazine Essay
3. Term “spectacular narcissism”	Time Magazine Essay
4. Stylistic effectiveness	Time Magazine Essay
5. Purpose of quote	Time Magazine Essay
6. Quote manipulates word patterns	Time Magazine Essay
7. Purpose of metaphors	Time Magazine Essay
8. Implications of metaphors	Time Magazine Essay
9. Use of figure of speech	Time Magazine Essay
10. Purpose of specific phrases	Time Magazine Essay
11. Character’s final response	Yeats Play
12. Character rejects food and money	Yeats Play
13. Character’s growing interest	Yeats Play
14. Example of literacy device	Yeats Play
15. Character’s reaction	Yeats Play
16. Character reaction example of	Yeats Play
17. Irish temperament conveyed with	Yeats Play
18. Character’s failure to understand	Yeats Play
19. Interpret Character’s words	Yeats Play
20. Phrase reinforces idea of	Yeats Play
21. Playwright’s attitude	Yeats Play
22. Playwright promotes idea of	Yeats Play

## SIBTEST Results for the English 30 Reading Comprehension Items Using Confirmatory and Exploratory Purification Procedures

Confirmatory						Exploratory					
Matching			Studied			Matching			Studied		
Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$
P	11	0.078 *	E	1	0.013	E	1	0.006	E	3	0.094 *
P	12	0.024	E	2	0.013	E	2	0.005	E	6	-0.070 *
P	13	-0.028	E	3	0.095 *	E	4	0.016	E	9	-0.074 *
P	14	-0.027	E	4	0.023	E	5	-0.011	P	11	-0.082 *
P	15	0.037	E	5	0.001	E	7	0.018		<b>Total</b>	<b>0.132</b> **
P	16	0.021	E	6	0.058 *	E	8	-0.037			
P	17	-0.014	E	7	0.028	E	10	0.002			
P	18	0.044	E	8	-0.026	P	12	0.020			
P	19	-0.001	E	9	0.066 *	P	13	-0.040			
P	20	-0.041	E	10	0.012	P	14	-0.030			
P	21	-0.005		<b>Total<sup>a</sup></b>	<b>0.035</b>	P	15	0.027			
P	22	0.023				P	16	0.012			
						P	17	-0.019			
						P	18	0.031			
						P	19	-0.007			
						P	20	-0.037			
						P	21	-0.014			
						P	22	0.019			

\*  $|\hat{\beta}_{UNI}| \geq 0.05$  and  $p < 0.05$ . \*\*  $p < 0.05$ .

Note. The English 30 passage comprehension sections measures the essay (E) and play (P) dimensions. A positive  $\hat{\beta}_{UNI}$  favors males.

<sup>a</sup> Items measuring dimension 2 served as the matching subtest because this dimension contained a larger number of items and, therefore, it was expected to produce a more reliable matching subtest. Neither the play nor the essay dimensions were predicted to elicit gender differences.

Table 3.

*Items on the Test Anxiety Inventory*

Item	Dimension
1. Confidence and relaxed	--
2. Uneasy, upset feeling	Emotionality
3. Thinking about course grade	Worry
4. Freeze on important exams	Worry
5. Will I get through school	Worry
6. Harder I work, more confused I get	Worry
7. Thoughts about doing poorly	Worry
8. Jittery on important exams	Emotionality
9. Feel well-prepared but nervous	Emotionality
10. Uneasy about test results	Emotionality
11. Tense during test	Emotionality
12. Wish exams didn't bother me	--
13. Stomach upset during exam	--
14. Defeat myself during exam	Worry
15. Panicky during exam	Emotionality
16. Worry before exam	Emotionality
17. Think about consequences of failure	Worry
18. Heart beating fast during exam	Emotionality
19. Try to stop worrying but can't	--
20. So nervous, forget facts	Worry

Table 4.

*Poly-SIBTEST Results for the Test Anxiety Inventory Using Confirmatory and Exploratory Purification Procedures*

Confirmatory						Exploratory					
Matching			Studied			Matching			Studied		
Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$	Dimension	Item	$\hat{\beta}_{UNI}$
W	3	0.085	E	2	-0.024	--	1	0.011	W	5	0.280 *
W	4	-0.030	E	8	-0.096	E	2	0.018	W	7	0.131 *
W	5	0.138 *	E	9	-0.122 *	W	3	0.052	--	13	-0.081 *
W	6	-0.013	E	10	-0.131 *	W	4	-0.058	W	14	0.224 *
W	7	0.119 *	E	11	-0.068	W	6	-0.038	E	16	-0.148 *
W	14	0.145 *	E	15	-0.156 *	E	8	0.031	W	17	-0.179 *
W	17	0.194 *	E	16	-0.317 *	E	9	0.032		<b>Total</b>	<b>0.345 **</b>
W	20	0.013	E	18	-0.056	E	10	-0.068			
				<b>Total</b>	<b>-0.342 **</b>	E	11	0.060			
						--	12	0.004			
						E	15	-0.021			
						E	18	0.091			
						--	19	0.037			
						W	20	-0.020			

\*  $|\hat{\beta}_{UNI}| \geq 0.100$  and  $p < 0.05$ . \*\*  $p < 0.05$ .

Note. The TAI measures the emotionality (E) and worry (W) dimensions of test anxiety. A positive  $\hat{\beta}_{UNI}$  favors males.