The purpose of this book is to provide a perspective and foundation for DIF, a review DIF approaches. Further, DIF methods classified into two categories: IRT methods and non-IRT methods, respectively. For the former, the estimation of an IRT model is required, and a statistical testing procedure is followed, based on the asymptotic properties of statistics derived from the estimation results. For the latter, the detection of DIF items is usually based on statistical methods for categorical data, with the total test score as a matching criterion. Item Response Theory (IRT) techniques provide a powerful means of testing items for bias, using what is known as differential item functioning (DIF) analysis. In contrast, Classical Test Theory (CTT) based methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups. This book provides different studies to detect a gender related DIF in mathematics.

Nabeel Abedalaziz

Dr. Nabeel Abedalaziz is a senior Lecturer in the Department of Educational Psychology at the University of Malaya/Malaysia where he teaches courses in Educational Psychology and Research Methodology.
Nabeel Abedalaziz

Detecting Differential Item Functioning
Nabeel Abedalaziz

Detecting Differential Item Functioning

LAP LAMBERT Academic Publishing
Impressum / Imprint
Alle in diesem Buch genannten Marken und Produktamen unterliegen warentechnen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.
Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite / Printed at: see last page
ISBN: 978-3-659-71756-7

Copyright © 2015 OmniScriptum GmbH & Co. KG
Alle Rechte vorbehalten. / All rights reserved. Saarbrücken 2015
<table>
<thead>
<tr>
<th>Content</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter One: Differential Item Functioning and Bias</strong></td>
<td></td>
</tr>
<tr>
<td>- Introduction</td>
<td>3</td>
</tr>
<tr>
<td>- Test Bias</td>
<td>3</td>
</tr>
<tr>
<td>- Test Fairness</td>
<td>5</td>
</tr>
<tr>
<td>- Differential Item Functioning (DIF)</td>
<td>9</td>
</tr>
<tr>
<td>- Differential item functioning and Bias</td>
<td>11</td>
</tr>
<tr>
<td>- DIF Forms</td>
<td>12</td>
</tr>
<tr>
<td>- Classical Test Theory (CTT)</td>
<td>13</td>
</tr>
<tr>
<td>- Test Validity in CTT Termonology</td>
<td>16</td>
</tr>
<tr>
<td>- Laten Trait Theory (IRT)</td>
<td>19</td>
</tr>
<tr>
<td>- IRT Model</td>
<td>20</td>
</tr>
<tr>
<td>- Classical Test Theory versus Item Response Theory</td>
<td>30</td>
</tr>
<tr>
<td>- IRT assumptions</td>
<td>32</td>
</tr>
<tr>
<td>- Parameter Estimation (a, b, and c)</td>
<td>36</td>
</tr>
<tr>
<td><strong>Chapter Two: Approaches to Test and Test Bias/DIF</strong></td>
<td>38</td>
</tr>
<tr>
<td>Introduction</td>
<td>38</td>
</tr>
<tr>
<td>Detecting DIF Approaches</td>
<td>40</td>
</tr>
<tr>
<td>Non IRT Based Approaches</td>
<td>45</td>
</tr>
<tr>
<td>- Logistic Regression (LR)</td>
<td>45</td>
</tr>
<tr>
<td>- Mantel-Haenszel method (M-H)</td>
<td>51</td>
</tr>
<tr>
<td>- Root-mean weighted squared differences (Standardized p- difference)</td>
<td>55</td>
</tr>
<tr>
<td>- Chi Square Type DIF Methods</td>
<td>56</td>
</tr>
<tr>
<td>- Equal Regressions</td>
<td>58</td>
</tr>
<tr>
<td>- The Mean and Covariance Structure Model (MACS)</td>
<td>59</td>
</tr>
<tr>
<td>- Transformed Item Difficulty</td>
<td>59</td>
</tr>
<tr>
<td>- The Simultaneous Item Bias Test (SIBTEST)</td>
<td>61</td>
</tr>
<tr>
<td>- Exploratory factor analysis Approach (EFA)</td>
<td>62</td>
</tr>
<tr>
<td>- IRT Based Approaches</td>
<td>63</td>
</tr>
</tbody>
</table>
- b- Difference Index ................................................................. 63
- Area index ........................................................................... 65
- Area Index of Two-Parameter Logistic Model ...................... 66
- Area Index of Three-Parameter Logistic Model ...................... 67
- Root mean squared difference (RMSD) .................................. 68
- Probability difference Indices ............................................... 69
- The likelihood-ratio (LR) ...................................................... 70
- Detecting DIF using 3-Parameter Logistic Model ................... 71
- Rasch Model Approach ....................................................... 71
- Partial Credit Approach (PC) ............................................... 73
- References ........................................................................... 74

- Chapter Three: Detecting Gender Related DIF using Logistic Regression and Mantel-Haenszel Approaches .................................................. 83
- Chapter Four: Detecting DIF: Comparisons between CTT and IRT Methods ................................................................. 98
- Chapter Five: DIF in Cognitive Components of Solving Mathematical Routine Problems ................................................................. 113
- Chapter Six: Detecting DIF Using TID ................................... 134
- Chapter Seven: A Gender-Related Differential Item Functioning of Mathematics Test Items ................................................................. 143
- Chapter Eight: Detecting DIF Using Logistic Regression and Transformed Item Difficulty ................................................................. 154
- Chapter Nine: Detecting DIF using Item Characteristic Curve Approaches ................................................................. 165
- Chapter Ten: Detecting DIF using Logistic Regression ................................................................. 177
Chapter One

Differential Item Functioning and Bias

Introduction

For instrument design and item development, the validity of a measurement instrument depends on the quality of the items included in the instrument. Furthermore, the reliability of test scores can be compromised by random measurement error, and the validity of test score interpretations can be compromised by response biases that systematically obscure the psychological differences among respondents. Psychological tests are often used to make important decisions that affect the lives of real. To the degree that such decisions are based on tests that are biased in favour of or against specific group of people, such biases have extremely important personal and societal implications. As such, one of the important issues faced by counseling psychologists is that of responding to the diversity of clients. In particular, it is important that the tests used by psychologists or test users be free of systematic demographic subgroup bias. Test bias has become an increasingly important topic in educational and psychological measurement. Both moral and legal questions as to the suitability of tests for various groups (e.g. ethnicity and gender) have been raised. Investigation of these questions is made more difficult in that there is no one definition of test bias.

Bias comes in many forms. It can be sex, cultural, race, or religious bias. An item may be biased if it contains content that is differentially familiar to subgroups of examinees, or if the item structure is differentially difficult for subgroups of examinees. An example of content bias against girls would be
one in which students are asked to compare the weights of several objects, including a football. Since girls are less likely to have handled a football, they might find the item more difficult than boys, even though they have mastered the concept measured by the item (Scheuneman, 1982).

Educational tests are considered biased if a test design, or the way results are interpreted and used, systematically disadvantages certain groups of students over others. For example, female students tend to score lower than males (possibly because of gender bias in test design), even though female students tend to earn higher grades in college on average (which possibly suggests evidence of predictive-validity bias).

Identifying test bias requires that test developers and educators determine why one group of students tends to do better or worse than another group on a particular test. As student populations in public schools become more diverse, and tests assume more central roles in determining individual success or access to opportunities, the question of bias—and how to eliminate it—has grown in importance. “Investigation of item bias involve gathering empirical evidence concerning the relative performance on the test item of members of the minority group of interest and members of the group that represent majority----DIF rather than bias is used commonly to described the empirical evidence obtained in investigation of bias” (Hambleton, Swaminathan, & Rojers, 1991, p. 109). Measurement experts frequently investigate differential item functioning (DIF) for demographic groups to ensure that tests are fair. DIF studies can also be conducted as a way of checking the stability of the item properties across important subgroups. Regardless of the groups examined, DIF is considered a serious threat to test validity because it implies that one group has an unfair advantage on an item
in comparison with another group. Logically, the first step in detecting test bias is to locate examination items on which one group of test-takers performs significantly better than another group (Roever, 2005; Perdrajeta, 2009).

**Test Bias**

Prior to the decision making based on test scores, it is critical to avoid bias, which may unfairly influence examinees' scores. There are two methods used to detect test biases. Roughly speaking, the two types of test bias reflect biases in the meaning of a test and biases in the use of a test. Test bias is a psychometric concept embedded in theories of test score validity.

Test score bias is defined within psychometric theories along with specific statistical and research methods that can be used to scientifically evaluate test score bias hypotheses allowing researchers to make informed decisions regarding test bias. Bias is the presence of some characteristic of an item that results in differential performance for individuals of the same ability but from different ethnic, sex, cultural, or religious groups (Hambleton & Jane, 1995).

A biased test is one in which there are systematic differences in the meaning of test scores associated with group membership. Another way of saying this is that a biased test is one in which people from two groups (e.g. gender: male and female; race: Malay, Chinese and Indian) who have the same observed score do not have the same standing on the trait of interest. A third way of saying this is that using a test to predict some criterion of interest results in systematic over- or under- prediction based on group membership. The most intuitive definition of bias is observation of a mean difference between groups. However, the mean difference by itself is a bad choice of
models of bias. This is because a mean difference could demonstrate bias, but it could also reflect a real difference between groups. The related literature categorized test/item bias into four categories, namely: construct-validity bias, content-validity bias, and item-selection bias and bias in criterion-related validity.

For instance, construct-validity bias refers to whether a test accurately measures what it was designed or supposed to measure. Therefore, Construct-validity bias occurs when a test has different meanings for two groups, in terms of the precise construct that the test is intended to measure. Construct bias has to do with the relationship of observed scores to true scores on a test. If this relationship can be shown to be systematically different for different groups, then we might conclude that the test is biased. On an intelligence test, for example, students who are learning English will likely encounter words they haven’t learned, and consequently test results may reflect their relatively weak English-language skills rather than their intellectual abilities. Thus, construct-validity bias can lead to situations in which two groups have the same average true score on a psychological construct but different average test scores.

The content-validity bias occurs when the content of a test is comparatively more difficult for one group of students than for others. It can occur when members of a student subgroup, such as various minority groups, have not been given the same opportunity to learn the material being tested, when scoring is unfair to a group, or when questions are worded in ways that are unfamiliar to certain students because of linguistic or cultural differences. Item-selection bias refers to the use of individual test items that are more suited to one group’s language, cultural experiences or cultural bias.
Predictive-validity bias (or bias in criterion-related validity) refers to a test’s accuracy in predicting how well a certain student group will perform in the future. For example, a test would be considered “unbiased” if it predicted future academic and test performance equally well for all groups of students. For example, college admissions officers might use SAT test scores to predict freshman GPAs. The SAT would be the predictor test and GPAs would be the outcome measure. In this context, test bias concerns the extent to which the link between predictor test true scores and outcome test observed scores differ for two groups. If the SAT is more strongly predictive of GPA for one group than for another, then the SAT suffers from predictive bias, in terms of its use as a predictor of GPA.

Further, the following are several representative examples of other factors that can give rise to test bias: A test is not demographically or culturally representative of the students who will take the test, test items may reflect inadvertent bias; the test may be biased if the “norming process” does not include representative samples of all the tested subgroups; the choice of language in test questions can introduce bias; a test includes references to cultural details that are not familiar to particular student groups; a certain test formats may have an inherent bias toward some groups of students, at the expense of others. For example, evidence suggests that timed, multiple-choice tests may favor certain styles of thinking more characteristic of males than females, such as a willingness to risk guessing the right answer or questions that reflect black-and-white logic rather than nuanced logic. and tests may be considered biased if they include references to cultural details that are not familiar to particular student groups.
In general, item bias occurs when examinees of one group (e.g., males) are less likely to endorse an item than examinees of another group (e.g., females) because of some characteristic of the test item or testing situation that is not relevant to the test purpose (Zumbo, 1999).

DIF is required, but not sufficient, for item bias. According to Zumbo (1999), Item impact and item bias differ in terms of whether group differences are based on relevant or irrelevant properties (respectively) of the test. Item impact is evident when examinees from different groups (males and females) have differing probabilities of endorsing an item because there are true differences between the groups in the underlying ability being measured by the item. Adverse impact is a legal term describing the situation in which group differences in test performance result in disproportionate examinee selection or related decisions (e.g., promotion). This is not evidence for test bias (Zumbo, 1999, p. 12). Item impact can be measured through the proportion of test takers passing an item regardless of their total score.

Finally, Clauser and Mazor (1998) indicated that the entire domain of item bias is really about policy. According to Zumbo (1999), if bias is being “legislated” from an outside body, this legislation will help you determine the answer to the following five operational policy matters of essential concern:

1. There are different sub-groups one could contrast. As such, there is a need to be clear as to which ones are of personal and moral focus.
2. You need to discuss how much DIF does one need to see before one puts the item under review or study.
3. Should an item only be sent for review by experts if it is flagged as favouring the reference group? Or should an item be sent for review irrespective of whether it favours the reference or focal group?
4. The suitable timing of the DIF analysis is also important. For instance, when you have a ready-made test that you are adopting for use, DIF analyses should be performed at the pilot testing stage or before scores are reported. In another hand, when you are developing a new test, DIF analyses should be conducted at pilot testing and certainly before any norms or cut-off scores are established.

5. What does one do when one concludes that an item is revealing DIF? Does one immediately dispense with the item or does one sends it to content experts for further validation studies?

**Test Fairness**

“The most highly charged issue surrounding testing, and the one of greatest importance to the public, is that of test fairness” (Hambleton, Swaminathan, & Rojers, 1999, p. 109). Fairness has to do with how a test is used. A judgment of fairness rests on values and reasonable people may disagree about the fairness of a test when both agree about the facts of the matter. Fairness is a social rather than a psychometric concept. The Standards notes four possible meanings of “fairness.” The first meaning views fairness as requiring equal group outcomes (e.g., equal passing rates for subgroups of interest). The second meaning views fairness in terms of the equitable treatment of all examinees (i.e., equitable treatment in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration). The third meaning views fairness as requiring that examinees have a comparable opportunity to learn the subject matter covered by the test. The fourth meaning views fairness as a lack of predictive bias.
In preparing an item bias review form, each item can be evaluated from two perspectives: Is the item fair? Is the item biased? While the difference may seem trivial, some researchers contend that judges cannot detect bias in an item, but can evaluate or assess an item's fairness (Hambleton & Jane, 1995). It is important to distinguish test fairness from test score bias. Test score bias is defined within psychometric theories along with specific statistical and research methods that can be used to scientifically evaluate test score bias hypotheses allowing researchers to make informed decisions regarding test bias. Test fairness, on the other hand, does not refer to a psychometric property of a test. Test fairness has to do with the appropriate use of test scores and is a social or philosophical or perhaps a legal term that represents someone’s value judgment (Thorndike, 1971).

Given the fact that test results continue to be widely used when making important decisions about students, the following strategies have identified that can reduce, if not eliminate, test bias and unfairness:

1. Having test materials reviewed by experts trained in identifying item and test bias.
2. Ensuring that norming processes and sample sizes used to develop norm-referenced tests are inclusive of diverse student subgroups and large enough to constitute a representative sample.
3. Screening for and eliminating items, references, and terms that are more likely to be offensive to certain groups.
4. Using multiple assessment measures to determine academic achievement and progress, and avoiding the use of test scores, in exclusion of other information, to make important decisions about students.
Differential Item Functioning (DIF)

There exists several definitions of DIF. For instance, “DIF refers to differences in item functioning after groups have been matched with respect to ability or attribute that the item purportedly measures... DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to attribute measured by the item and the test on which it appears” (Dorans & Holland, 1993, P. 37). Angoff (1993) points out that an item which displays DIF has different statistical properties in different group settings when controlling for differences in the abilities of the groups. It is important to stress that DIF is an unexpected difference between two groups after matching on the underlying ability that the item is intended to measure. According to Hambleton, Swaminathan, and Rojers (1999), “an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right” (p. 110). As such, the DIF analysis of one specific test item is as independent as possible from the DIF analyses of the other test items (Zumbo, 1999).

DIF is a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of different groups of examinees. As such, DIF aids in the identification of test items that are potentially biased. In assessing response patterns, the comparison groups are initially matched on the underlying construct of interest. By matching groups on the measured variable, test developers are better able to determine whether item responses are equally valid for distinct groups of test-takers (Zumbo, 1999). DIF methods, therefore, assess the test-takers’ response patterns to specific test items.
Differential item functioning and Bias

To reiterate, a test item is considered to be biased when a dimension on the test is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking a test (Hambleton, Swamanithan & Rogers, 1991). Thus, if DIF is not evident for an item, then there is no item bias. Conversely, DIF is required but is not sufficient for item bias. That is, if DIF is apparent, then its presence is not sufficient to declare item bias; rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical evaluation) to determine the presence of item bias. According to Clauser and Mazor (1998), it is vital to distinguish between item bias and DIF from inappropriate item content or framing that is potentially offensive. Furthermore, if an item is offensive to the all of test-takers, it is not going to be detected as biased.

An item might show DIF, but not be considered biased if the difference is a result of the actual difference in the groups’ ability to respond to the item. If test-takers differed in knowledge, a difference in item responses would be expected. Consequently, a difference in the performance of groups of examinees with different abilities on specific items is not indicative of test bias, but rather of item impact (Schumacher, 2005). But it can be added, that in order to be able to determine whether an item that shows DIF is biased or not, further analysis have to been done (Camilli & shepherd, 1994). It is then of interest to determine whether the differences deepened on differences of ability of the compared groups (not biased) or on the item measuring something else than intended (biased). DIF is a statistical property of an item
while item bias is more general and lies in the interpretation (Camilli & Shepard, 1994).

**DIF Forms**

Two distinct forms of DIF have been recognized. These have been called uniform and non-uniform DIF. In IRT terminology, uniform DIF, occurs when two ICC’s differed, but are more or less parallel (Hambleton, Swamanithan & Rogers, 1991). Furthermore, uniform DIF is likely to occur when two ICC’s have different b (difficulty) parameters and similar a (discrimination or slope) parameters (Swaminathan & Rogers, 1990). Figure 1 illustrates a clear case in which the two ICCs, one for focal group (females) and one for reference group (males), differed, but are more or less parallel. In this case, female examinees do poorer than the male examinees at all ability levels. Therefore, this item reveals a uniform DIF in the sense of operating the same way in different groups and produces an overall effect on the mean differences between the two groups. Whereas, Figure 2 illustrates a clear case in which the two ICCs, one for focal group (females) and one for reference group (males), cross at the ability of -.5. In this case, female examinees with low ability levels do poorer than the male examinees. The opposite is true for the examinees at the high ability levels. Therefore, this item reveals nonuniform DIF in the sense of not operating the same way in different groups but does not produce an overall effect on the mean differences between the two groups.
Non uniform DIF occurs when there is an interaction between ability level and subgroup membership (Swaminathan & Rogers, 1990), and the
result is that the ICC’s for the two subgroups cross at some ability value (Hambleton, Swamanithan & Rogers, 1991). Before the crossover point, the item is favoring one subgroup, and after the ICC’s cross, the item starts to favor the other group, so the DIF could cancel them out, and the item shows no net DIF. Thus, non uniform DIF is likely to occur when the two ICC’s have similar b parameters and different maximum slopes. In general, Uniform DIF is said to apply when differences between groups in item responses are found at all ability levels, while in non-uniform DIF an interaction is found between ability level, group assignment, and item responses (Harvey & Greenberg, 1996; Sells, 1973).

According to Mellenbergh (1982), uniform DIF occurs when there is no interaction between ability level and group membership. Non uniform DIF occurs when there is interaction between ability level and group membership. Swaminathan & Rogers (1990) distinguished two types of nonuniform DIF. Ability typically falls in the range -3 to +3 on the ability level scale in item response theory. When the ICCs cross in the middle of this range, a type of non uniform DIF occurs that is analogous to a disordinal interaction in analysis of variance (ANOVA) models (Swaminathan & Rogers, 1990). When the ICCs cross outside this range or when the ICCs are not parallel but do not cross, a type of nonuniform DIF analogous to ordinal interaction occurs. Li & Stout (1993) termed these two types of DIF nondirectional and unidirectional, respectively.

DIF is basically found by examining differences in ICCs across groups, where as Differential Test Functioning, or DTF, is the analogous procedure for determining differences in Test Characteristics Curves, or TCCs. Further, DTF is arguably more important because it speaks to impact (DIF in one item
may be significant, but might not have too much practical impact on test results).

**Classical Test Theory (CTT)**

From the perspective of classical test theory, an examinee's obtained test score \(X\) is composed of two components, a true score component \(T\) and an error component \(E\):

\[ X = T + E \]

The true score component reflects the examinee's status with regard to the attribute that is measured by the test, while the error component represents measurement error.

Measurement error is randomized error. It is due to factors that are irrelevant to what is being measured by the test and that have an unpredictable (unsystematic) effect on an examinee's test score.

The score you obtain on a test is likely to be due both to the knowledge you have about the topics addressed by exam items \(T\) and the effects of random factors \(E\) such as the way test items are written, any alterations in anxiety, attention, or motivation you experience while taking the test, and the accuracy of your "educated guesses."

Whenever we administer a test to examinees, we would like to know how much of their scores reflect "truth" and how much reflects error. It is a measure of reliability that provides us with an estimate of the proportion of variability in examinees' obtained scores that is due to true differences among examinees on the attribute(s) measured by the test.
When a test is reliable, it provides dependable, consistent results and, for this reason, the term consistency is often given as a synonym for reliability (e.g., Anastasi, 1988).

Ideally, a test’s reliability would be calculated by dividing true score variance by the obtained (total) variance to derive a reliability index. This index would indicate the proportion of observed variability in test scores that reflects true score variability.

In summary, we can sum up the main features of CTT as follows:

1. CTT analysis is the easiest and most widely used form of analysis. The statistics can be computed by generic statistical packages (or at a push by hand) and need no specialist software.

2. CTT analysis is performed on the scale as a whole rather than on the item and although item statistics can be generated, they apply only to that group of students on that collection of items.

3. CTT assumes that any test score is comprised of a “true” value, plus randomized measurement error.

4. CTT assumes that this randomized measurement error is normally distributed; uncorrelated with true score and the mean of the error is zero.

\[ X_{obs} = R_{true} + G(0, \sigma_{err}) \]

5. In CTT analysis, item difficulty is the proportion of examinees who answered the question incorrectly. For essay item, it is the average score expressed as a proportion. Given on a scale of 0-1, the higher
the proportion the greater the difficulty. In another words, the p-value is relative to the ability level of examinees

6. The discrimination of an item (d-value) is the (Pearson) correlation between the average item score and the average total test score. Being a correlation it can vary from −1 to +1 with higher values indicating (desirable) high discrimination. The d-value is relative to the homogeneity of the ability levels of the examinees. The subject-matter homogeneity of the test items, and the variation (dispersion) of p-values of test items.

7. Reliability is a measure of how well the test or survey “holds together”. For practical reasons, internal consistency estimates are the easiest to obtain which indicate the extent to which each item correlates with every other item. This is measured on a scale of 0-1. The greater the number the higher the reliability. Further, the reliability is relative to the standard deviation of the test scores, and to the items difficulty (p-values) and items discrimination (d-values), all of which are dependent on all of which are dependent upon particular abilities of the examinees and the characteristics of the test (Warm, 1978).

8. The mean scores, variance, p-values, d-values, reliability, skewness and kurtosis are relative to the characteristics of the tests and of the examinees.

9. According to Warm (1977), CTT statistics are meaningful only in an extremely limited situation. For instance, when the same item is given to the same population as part of strictly parallel tests. Such a situation is rarely occurs. Furthermore, the basic precepts or
concepts and definitions of CTT are untestable. They are taken as true without any way to provide empirical evidence of their relevance to the reality (tautologies). Thus, no one knows if the CTT models applies to any real test.

Test Validity in CTT Termonology

According to Zumbo (1999), traditional view of validity focuses on: whether a scale is measuring what we suppose it is, test reliability as a necessary but not sufficient condition for validity, validity as a property of the measurement instrument, validity as defined by a set of statistical techniques, a measure is either valid or invalid, and various types of validity (content, concurrent, predictive, and construct). Whereas, the current view of validity focuses on: construct validity is the vital feature of validity; there is debate as to whether reliability is a necessary but not sufficient condition for validity; this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in ones inferences; validity is no longer a property of the test or instrument but rather of the inferences made from that test; the validity conclusion is on a continuum and not simply declared as valid or invalid; validity is no longer defined by a set of statistical techniques with a gold-standard but rather by an elaborated theory and supporting methods; consequences of test decisions and use are an essential part of validation; and there are no longer various types of validity so that it is no longer acceptable in common practice that the test user or researcher assumes that he/she only needs to demonstrate one of the four types to have validity (Zumbo, 1999, P. 9-10).
Laten Trait Theory (IRT)

Item response theory (IRT) is a modern measurement approach that relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of providing a particular response (Stark, Chernyshenko, Lancaster, Drasgow, & Fitzgerald, 2002). IRT has many advantages over Classical test theory (CTT), with the major distinction relating to the property of invariance. Whereas CTT statistics depend on the items included in the test and persons examined, IRT item and person parameters are invariant depending neither on the subset of items used, nor on the distribution of the latent trait in the population of respondents (Stark, Chernyshenko, Chuah, Lee, & Wadlington, 2001).

The most salient feature of IRT that distinguishes it from CTT is the property of invariance of item and ability parameters (Hambleton, Swaminathan & Rogers, 1991). Invariance means that item parameters (e.g., difficulty, discrimination and guessing) are not dependent on the ability distribution of any particular group of examinees and the examinee ability parameter ($\theta$) is not dependent on a specific set of test items. This also implies that for a correctly specified IRT model, the ICC for two subpopulations of examinees will be the same regardless of the groups’ ability distributions (Hambleton, Swaminathan & Rogers, 1991). This property makes IRT an attractive framework for examining DIF since the occurrence of non-coinciding ICCs is an indicator of differential item functioning between two groups.

In summary, IRT is useful in overcoming the weaknesses of classical test theory for the following reasons. IRT is not item or sample-dependent.
That is, the assumption of local independence and the property of invariance allow item parameters to be estimated for any group of examinees and permits estimations of examinee ability that are not test-dependent. Also, IRT provides a mathematical model for predicting item success, conditional on ability level.

Finally, test reliability in IRT is based only on the items chosen for a test rather than on a specific group of test takers and the availability of parallel forms such as are required for classical test theory. Moreover, the assumptions of IRT are explicit and have the potential of empirical testing. It is possible to discover if the data reasonably meet IRT assumptions. IRT also provides an extremely powerful tool for special studies such as: item bias and DIF, ordered polytomous items, scaling individuals for further analysis, test construction and modification, computer adaptive testing, and test equating.

When we give several tests of science (or any subject) to a group of examinees, we will find that the same examinees score high on the test and the same examinees score low. In other word, we will find a consistency in examinees performance (ability) on the different science tests. This consisyency is due to something called “mental trait”. As such, mental trait is the characteristics of the examinees that causes consistent performance on the tests, whatever, if anything, it is. A trait is never observed directly, therefore, it is called a latent “trait” (Warm, 1978).

In IRT, a latent trait is symbolized as theta (θ) which refers to a statistical construct and measured by “logit unit”. In cognitive tests, latent traits are called the ability measured by the test. The total score on a test is taken as an estimate of that ability. Theta is a continuum from minus infinity (-∞) to plus infinity (+∞), it has no absolute zero point or unit. Therefore, the zero point and unit are often taken as the mean and standard deviation,
respectively, of some reference sample of examinees. Mathematically, the theta values belong to the interval [-3, +3]. Further, the theata values of a sample need not be distributed normally.

The normal curve (a bell-shaped) is called normal frequency function and given by the equation:

$$ N(0,1) \equiv y = \frac{e^{-\frac{Z^2}{2}}}{\sqrt{2\pi}} $$

where $Z$ is a standardize score.

The Logistic frequency function has a mean equal zero and standard deviation approach to one, and given by the equation:

$$ L(0,D) \equiv y = \frac{De^{-1.7Z}}{(1+e^{-1.7Z})^2} $$

Where:

$Z$: is a standardize score.

e is a transcendenal number whose value is approach to 2.72 (the base of natural logarithm).

$D$: is chosen to allow the logistic frequency function to approximate the normal frequency function as closely as possible ($D$ approach to 1.7).

The normal cumulative frequency curve (S-shaped curve) is called an “ogive”. This curve gives the proportion of area under the normal curve (A bell-shaped curve) that lies to the left of each point on abscissa. We call an ogive curve as a normal distribution function or “normal ogive”. Further, the logistic ogive (logistic distribution function) may be taken as close approximation to the normal ogive model (item characteristics curves: ICCs or item response functions: IRFs). There is an infinite family of ICCs, each
different in some way from every each other (Warm, 1978). The logistic ogive (ICC) is given by the equation:

\[
\int_{-\infty}^{Z} L(O, 1.7) = \frac{1}{1+e^{-1.7Z}}
\]

ICC is the regression of observed responses to an item, on unobservable traits, underlying the performance of the group of examinees. (Figure 3). This nonlinear regression function represents the probability of an examinee, at a given ability level, answering the item correctly. The item characteristics curves (ICCs) are strictly monotonic functions. Because they are going from left to right, the ICC always gets higher and higher, never is completely horizontal, and never goes down. If the item discrimination is greater than zero. In this case, an examinee with high level ability should have more chance of endorsing the item than one with the lowest ability level. In another words, the principal conceptual unit of IRT is the item characteristic curve (ICC). An ICC is the function that relates the probability of a correct answer on an item to the “ability” measured by the test containing the item. If the unidimensional assumption of the test is met, an item response function (IRF) or item characteristic curve (ICC) defined by its item parameters will remain unchanged across subpopulation groups. An ICC estimated from any group will be equal to an ICC from another, and both will be equal to the ICC estimated from responses of all examinees.
Three parameter logistic ogive may differ from each other in only three ways, one for each parameter. One way in which ICC may differ is in “b-parameter”. An item’s b parameter (difficulty) is the point on the ability scale corresponding to the location on the S-function item characteristic curve (inflection point) where the probability of a correct response is .5 (Hambleton et al., 1991). The inflection point is always the point where the slope is at its maximum. Thus, b-parameter is the horizontal position of the inflection point. In other words, b is determined by first locating of the point on the ICC that corresponds to a 50% chance of getting the item right (.5: the probability of a correct response (PCR) on the y-axis), and then determining the value of \( \theta \) (on the x-axis) that corresponds to that point on the ICC.

Items that are difficult will have higher b values and will be located at the right or higher end of the \( \theta \) scale, which indicates that a greater level of ability is required in order to answer them correctly. Conversely, easy items will have lower b parameter values, will stay to the left (lower) end of the \( \theta \)
scale, and will require less ability to answer them correctly (Hambleton et al., 1991). Typical b-values range from -2.5 to +2.5. For instance, a b-value of -2.5 indicates the item is very easy, whereas, an item with a =2.5 b-value is very difficult, and item with a 0.0 b-value is of medium difficulty.

The logistic ogive (ICC) has lower asymptote and upper asymptote. They are horizontal lines that the ogive approaches at its extremes, but never quite reaches. The upper asymptote is located on the vertical axis (y-axis) at \( y=1.0 \), but the upper asymptote cannot touch the line \( y=1 \) (the horizontal line). With respect to the lower asymptote, the lower part of the logistic ogive never quite reaches the lower asymptote. This lower asymptote is called pseudo-guessing parameter (c-parameter). The effect of c-parameter is to squeeze the ogive into a smaller vertical range (reduced change). Thus, the c-parameter reduces the slope of the ogive at every point on the \( \theta \) scale. Moreover, the inflection point on the logistic ogive is always half-way between its upper and lower asymptotes.

The c-parameter indicates the probability of examinees with very low ability of endorsing the item correct. Item with c-values of .30 or greater are not very good items, whereas, the c-values of .20 or less are desirable. The lower the c-values, the better, whereas, a zero c-value is ideal. Typically, the c-value is about \( (1/A - .05) \), where \( A \) is the number of alternatives. Thus, the c-values for 4-choice item is approach to .20, whereas, the c-value for 5-choice item is about .15 (Warm, 1978).

The item discrimination parameter (a-parameter) is related to the slope of the either ogive at the inflection point or in other words at the b-parameter. According to Warm (1977), for the logistic ogive with c =0.0, a-parameter can be calculated as follows: \( a=\sqrt{2\pi} m \), where \( m \) is the slope of the ogive at
the b-value. Figure 1 gives an example of a parametric ICC. The horizontal axis is the continuum of variation for the latent variable. The scale of the latent variable is in z scores. The item depicted in Figure 4 has discrimination (i.e., slope) of 1.055, difficulty (i.e., threshold) of 0.648, and guessing parameter of 0.255. Note that the item discrimination parameter determines how rapidly the curve rises from its lowest value of c, in this case 0.10, to 1.0. Note that if the curve is relatively flat then the item does not discriminate among individuals with high, moderate, or low total scores on the measure. Item discrimination values of 1.0 or greater are considered very well. Finally, note that the threshold parameter is the latent variable value on the continuum of variation at which the curve is midway between the lowest value, c, and .255, and therefore for achievement or aptitude measures, is a marker of the item difficulty. Items with difficulty values less than -1.0 indicate a fairly easy item whereas items with difficulty greater than 1.0 indicate rather difficult items.
The larger the a-parameter, the steeper the logistic ogive. The values of a-values range from 0.0 to ∞. An item with a low a-value discriminates poorly over a wide range of θ. With a high a-value the item discriminate well, but over a small range of θ. For most purposes, item with a-value below .80 is not very good item. Further, the a-parameter can be extracted as follows:

\[(p^{-1}(θ) - b) = 1/a\]

Where:

\(p^{-1}(θ)\): is the point on θ, where the hiegh of the ogive equal to: c+.8455(1-c).

The number (.8455) is the proportion of the area unde logistic frequency function, and to the left of z-score equal to 1.

a: is item discrimination.
b: is item difficulty.
Figure 5: Two hypothetical ICCs with different item discrimination and the same difficulty estimates
The ICCs shown in Figure 5 portray two items with equal item difficulty among respondents but different slope (i.e., equal discrimination among respondents). The slope of the curve indicates how well the item discriminates among levels of ability (a flat curve, for example, would mean that irrespective of the level of ability, individuals have the same likelihood of getting an item correct). As we can see, the item 2 has a low level of item discrimination than item 2.

The ICCs shown in Figure 6 portray two items with equal slope (i.e., equal discrimination among respondents) but different placements on the continuum of variation (item difficulties). One would need to have more of the latent variable to endorse the item depicted by the dashed line than by the solid line. The dashed line is further to the right on the continuum. The dashed line, item 2, is thus considered more difficult than item 1.
The one-parameter logistic model, explains the relationship between levels of the ability and probability of a correct response on the item in terms of the difficulty of the item. An item’s b parameter (difficulty) is the point on the ability scale corresponding to the location on the ICC where the probability of a correct response is 0.5 (Hambleton et al., 1991). Substituting θ for Z in the logistic ogive and subtracting the b-value, gives the item characteristic curves for the one-parameter logistic model by the equation:

\[ P_i(\theta) = \frac{e^{(\theta-b_i)}}{1 + e^{(\theta-b_i)}} \]

Where:

- \( P_i(\theta) \) : is the probability of a correct answer to the item (i), i=1, 2, --------, n.
- \( b_i \) : is the item i difficulty parameter.
- n: is the number of test items, and
- e: is a transcendental number whose value is 2.72.
The two-parameter logistic model makes use of the b parameter (item difficulty) just as in the one-parameter model, but adds an additional element which indicates how well an item separates examinees into different ability levels. The a parameter used in the two-parameter model is called the item discrimination parameter and is equal to the slope of the ICC when it is at its steepest (Hambleton et al., 1991). The normal ogive curves (ICCs) for the 2-parameter logistic model are given by the formula:

\[ p_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \]

Where:
- \( p_i(\theta) \): is the probability of a correct answer to the item (i), i=1, 2, ..., n.
- \( b_i \): is the item i difficulty parameter.
- \( a_i \): is the item i discrimination parameter.
- \( n \): is the number of test items, and
- \( e \): is a transcendental number whose value is 2.72.

The three-parameter logistic model builds upon the two-parameter model by adding pseudo-chance-level parameter c. The c parameter is the value of the lower asymptote of the item characteristic curve and is indicative of the probability that an examinee with a very low ability score would answer an item correctly. The normal ogive curves (ICCs) for the 3-parameter logistic model are given by the formula:

\[ p_i(\theta) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \]

Where:
- \( p_i(\theta) \): is the probability of a correct answer to the item (i), i=1,2, ..., n.
- \( b_i \): is the item i difficulty parameter.
- \( a_i \): is the item i discrimination parameter.
ci: is the item i pseudo-guessing
n: is the number of test items, and
e: is a transcendental number whose value is 2.72.

**Classical Test Theory versus Item Response Theory**

Although CTT has served the measurement community for most of this century, IRT has witnessed an exponential growth in recent decades. The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory). Although CTT’s major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model. At the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee’s ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinees on an item, well known as the p value of the item, is used as the index for the item difficulty (actually, it is an inverse indicator of item difficulty, with higher value indicating an easier item). The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination, which is often expressed statistically as the Pearson product-moment correlation coefficient between the scores on the item (e.g., 0 and 1 on an item scored right-wrong) and the scores on the total test. When an item is dichotomously scored, this estimate is often computed as a point-biserial correlation coefficient.
The major limitation of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT’s application in some measurement situations (e.g., test equating, computerized adaptive testing). Despite the theoretical weakness of CTT in terms of its circular dependency of item and person statistics, measurement experts have worked out practical solutions within the framework of CTT for some otherwise difficult measurement problems. For example, test equating can be accomplished empirically within the CTT framework (e.g., equipercentile equating). Similarly, empirical approaches have been proposed to accomplish item-invariant measurement (e.g., Thurstone absolute scaling) (Englehard, 1990). It is fair to say that, to a great extent, although there are some issues that may not have been addressed theoretically within the CTT framework, many have been addressed through ad hoc empirical procedures. IRT, on the other hand, is more theory grounded and models the probabilistic distribution of examinees’ success at the item level. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT’s primary focus on test-level information. The IRT framework encompasses a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models, known as three-, two-, and one-parameter IRT models.

Although tests have always been composed of multiple items, item response theory (IRT) takes a much more item-level focus than classical test
theory (CTT), which tends to focus more on test-level indices of performance (e.g., the overall reliability coefficient, or standard error of a scale). In particular, the focus on estimating an ICC for each item provides an integrative, holistic view of the performance of each item that is not readily available when using CTT-based methods to develop or examine a test. That is, although CTT can quantify the total-sample difficulty (e.g., as a p value) or discrimination (e.g., as an item-total biserial correlation) for an item, it lacks an effective means for simultaneously combining and presenting this information (including the role of guessing, or other factors that might lead to a nonzero lower asymptote) in an easily-used format.

With respect to test scoring, IRT-based tests – especially those based on the 2- or 3- parameter models – offer considerable advantages over the “number right” scoring methods typically used in CTT-based tests. Specifically, when estimating an examinee’s score using IRT, we can simultaneously consider the following sources of information: (a) which items were answered correctly/incorrectly (or in the keyed vs. non-keyed direction); and (b) for each of those items, the difficulty, discrimination, and nonzero lower-asymptote parameters of the item. This offers the potential to produce better estimates of the \( \theta \) scores, to produce quantitative estimates of the “quality” or likelihood of any given observed response profile (termed appropriateness indices), and to assess the degree to which the given IRT model provides a good “fit” to the pattern of responses produced by the individual in question.

Classical Test Theory (CTT) test score state that an examinee’s observed score consists of his/her true score plus error. IRT has a similar interest in determining an examinee’s true score (latent trait score). However,
CTT approaches are limited in that examinee ability is defined in terms of a particular test, and the difficulty of that test is determined by the ability of the examinees who take it. This circularity of item and examinee characteristics in CTT branches into the estimation of reliability and validity as well because the test and item characteristics change as the examinee pool changes. Item Response Theory models, contrary to CTT models, are falsifiable in that they may or may not be appropriate for a particular data set (Hambleton et al., 1991). IRT models do not suffer from the limitations of CTT because item and ability parameters are invariant under a linear transformation (i.e., it is possible to change the means and variance estimates for different subgroups so that they lie on the same metric). Estimates of item parameters obtained from different examinee groups will be the same, and estimates of examinee ability do not depend on the pool of items administered (except for sampling or measurement errors (Hambleton et al., 1991).

Lastly, Classical Test Theory is limited in that it can only provide test level information. There is no consideration of how examinees perform on individual items (other than via statistics such as the item p value). It is sometimes essential to be able to design tests with items targeted toward specific ability levels. IRT models allow a test developer to design items that, for example, discriminate well among high ability examinees (Hambleton et al., 1991). In short, IRT models, because they provide item level information, are far superior to CTT models for many testing applications, especially those that seek to examine the performance of individual test items.
IRT assumptions

There are two primary postulates of IRT: (a) Examinee performance on a test item is a function of latent traits, or abilities; and (b) the graphical relation between examinees’ latent traits and their probabilities of answering an item correctly is in the form of a monotonically increasing function called an item characteristic curve (ICC). In other words, item performance depends on latent traits (e.g., ability), and as the level of the latent trait increases, the probability of a correct response either increases or stays the same (Hambleton et al., 1991). In IRT models, the underlying latent trait is referred to as theta ($\theta$), which is conceptually similar to a “true score” in Classical Test Theory.

The graph of an item characteristic curve has, on its x-axis, $\theta$ (expressed typically as a Z-score ranging from -3 to +3), and on its y-axis, the probability of a correct response (PCR). There are also several assumptions about the data to which IRT models are applied. The first assumption is that of unidimensionality, that one ability (latent trait) is measured by a test. In order for this assumption to be adequately met in an IRT model, a set of test data must consist of a “dominant” factor from which overall test performance results (Hambleton et al., 1991). Local independence, while related to unidimensionality, is the assumption that when ability (the latent trait) is held constant, there should be no relation between examinees’ responses to different items (Hambleton et al., 1991). In other words, the underlying latent trait in which the test purports to measure should be the only factor that has an overall influence on responses to test items, and when that latent trait is statistically controlled, there should be nothing consistently affecting item performance, and thus, the items should be uncorrelated (independent).
When there is an adequate fit between an IRT model and a set of test data, there are several desirable results such as test-free measurement. Test-free measurement implies that the estimates of examinee ability are assumed to be the same even if a different set of items is used (barring measurement errors), and item parameter estimates will be identical for different groups of examinees (except for sampling errors (Hambleton et al., 1991). This property of invariance of item and ability parameters is one of the advantages of IRT models.

**Parameter Estimation (a, b, and c)**

Before being used (in an item bank or for measurement) items must first be calibrated. That is their parameters must be estimated. There are two main procedures - Joint Maximal Likelihood and Marginal Maximal Likelihood. JML is most common for one parameter logistic model and two parameters logistic model, while MML is used more frequently for three parameters logistic model. Bayesian estimation and estimated bounds may be imposed on the data to avoid high discrimination items being overvalued.
Chapter Two

Approaches to Test and Test Bias/DIF

Introduction

There are two approaches for examining potential measurement bias, namely: (a) judgmental, and (b) statistical (Zumbo, 1999). Approaches to test bias can be classified into two categories. The first category includes the approaches which rely upon the existence of a criterion external to the test through which the bias of the test can be examined. In the second category, no external criterion is used and bias is detected from the characteristics of test and test items. The former commonly focuses on predictive studies investigating criterion-scale relationships such as those in some personnel selection studies. External evidence of test bias is suggested when the relationship between the scores and criterion is different for the various groups. Internal evidence comes from the internal relationships of the items to each other. DIF is clearly a matter of internal item relationships (Zumbo, 1999). Approaches to test bias in the absence of a criterion have arisen from construct validation procedures.

Construct validation involves consideration of the internal structure of the test parts and the extent to which the observed structure is consistent with theoretically expected structure. The extension of this approach to bias consideration involves the extent to which the construct being assessed is comparable across different groups. This general notion can be applied using factor analysis. Based on factor analysis procedures, if the factor structure of the test is different across different groups, the test is shown to be unfair. In
other hand, if the factor structure of the test is similar in different groups, then this test provides no evidence of unfairness. This approach represents the first level of detecting test bias (i.e. test level).

The researchers have been turned to detect items bias. Several approaches have been proposed and applied to detect item bias in the absence of a criterion. For instance, some approaches are based on the stability of item difficulty across different groups. According to these approaches, the existence of variation in item difficulty across different groups provides an evidence for possible item bias. Based on ANOVA, the item by group interaction term has been used to detect item bias. The existence of the significant interaction is indicative of item bias. Another approach tried to examine the similarity of the rank order of the item difficulties (p-values) or transformed item difficulties.

Unfortunately, Lord (1980) notes that it appears to be unjustified to expect the plot of p-values to form a straight line (the best line of fit). Furthermore, the p-values and transformed item difficulties (Deltas) are group dependent and not directly comparable across groups. For cultural bias, classical parameters will generally not be linearly related across subgroups of population. This means that the test for cultural bias using classical parameters can lead to artifactual detection of bias. In addition, an interaction between item and groups in ANOVA may be the result of comparing different groups from different level of ability scale in a highly discriminating item even when the examinees groups have identical probabilities of correct response at comparable ability level. Accordingly, Lord recommended using IRT approaches for item bias detection. These approaches have the advantages over the CTT approaches (item analysis and factor analysis procedures) of
estimation item parameters which are independent of the ability level of group tested.

**Detecting DIF Approaches**

A plausible but not exhaustive classification of DIF detection techniques is as follows: Classical Test Theory CTT-based methods, Factor Analysis (FA-based methods), $\chi^2$-based methods, and Item Response Theory IRT-based methods. According to Cole and Moss (1989), the work on DIF has focused upon the last two approaches, namely, those based upon $\chi^2$ and IRT. Unlike CTT-based methods, these last two approaches are conditional methods. In turn, $\chi^2$-based methods can be divided into four different groups:

1. The $\chi^2$-based methods in the strict sense, i.e., the $\chi^2$ correct (Scheuneman, 1979) and the $\chi^2$ full (Camilli, 1979).
2. The Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), a natural outgrowth of the former $\chi^2$ methods, which is widely used and easy to implement.
3. The Loglinear Models (LM) (Mellenbergh, 1982) to test the conditional independence of group membership and the score on the studied item given the matching variable.
4. The Logistic Regression (LR) procedure (Rogers & Swaminathan, 1993).

Further, DIF methods can be categorized in accordance to eight criteria. The first criterion is whether the methods are parametric or non-parametric, i.e. whether the model of the item is in focus or the data material is in focus. Non-parametric methods are not based on a specific statistical model although they may rely on strong assumptions. These methods are particular useful when the sample sizes are small for the groups of interest (Camilli,
Parametric methods use a specified model to examine DIF. The second criterion is whether the matching variable is based on an observed (e.g. total test score) or a latent variable. In DIF analysis, matching examinees having same ability from reference and focal groups is an important issue. Because an item functioning differentially across groups is defined as an item affected by additional dimensions to that of specified by the matching criterion, matching criterion should be an adequate representation of all the dimensions required to respond an item correctly. Univariate and multivariate matching criterion can be used to specify the students of the same ability. In the univariate analysis, students matched on the total test scores. On the other hand two different perspectives were used in the multivariate analysis. It was hypothesized that, by matching on a more refined criterion, fewer items would be displayed DIF, because item performance would be compared for groups of students whose ability levels are presumable more similar than those matched on total score alone.

The third criterion is whether the method can handle or be extended for use with polytomously scored items. The fourth criterion is between if a DIF methods can measure the effect size of DIF and test DIF, i.e. measure and/or test DIF. According to Zumbo (1999), two points are noteworthy at this juncture. First, the statistical test should accompanied by some measure of the magnitude of the effect. This is necessary because small sample sizes can hide interesting statistical effects, whereas large sample sizes can point to statistically significant effect where the effect is quite small and meaningless (Kirk, 1996). Second, researchers urge to report effect sizes for both statistically significant and for statistically non-significant results. Without an examination of effect size, trivial effects could be statistically significant when
the DIF test is based on a large sample size. The fifth criterion includes which kind of DIF the methods can handle; i.e. uniform and/or non-uniform DIF. Most non-parametric methods can only handle uniform DIF satisfactorily, whereas, All parametric methods can handle both kinds of DIF. The sixth and final criterion is whether the method can handle the cut-off score in a special way or not. The seventh criteria is the number of focal groups, and whether or not item purification is used.

According to Magis, Beland, Tuerlincks, and De Boek (2010), a practical issue when detecting DIF is that the presence of one or several DIF items may influence the results of tests for DIF in other items. Thus, some non-DIF items can wrongly be identified as DIF items, which indicates an unwanted increase of the Type I error of the approach. This is especially the case if some DIF items are included in the set of a priori non-DIF items. Such a priori non-DIF items are usually called anchor items. For non-IRT methods, this implies that the total test scores, which are used as proxies for ability levels, are influenced by the inclusion of DIF items. For IRT methods, the DIF items have an unwanted effect on the scaling of the item parameters used to obtain a metric (Magis et al., 2010). As such, there is a need to purify the matching criterion in the process of conducting the DIF analysis. That is, items that are identified as DIF are omitted, and the total score is recalculated. This re-calculated total score is used as the matching criterion for a second DIF analysis. Again, all items are assessed. This matching purification strategy has been shown to work empirically (Zumbo, 1999). Holland and Thayer (1988) note that when the purification strategy is used, the item under detection process must be included in the matching criterion even if it was identified as revealing DIF on initial screening and excluded from the criterion
for all other items. That is, the item under study should always be included in its own matching criterion score in order to reduce Type I errors. According to Mais et al (2010), item purification can be sketched by using the following stepwise process.

1. Test all items one by one, assuming they are not DIF items.
2. From results of Step 1, define a set of DIF items.
3. If the set of DIF items is empty after the first iteration, or if this set is identical to the one obtained in the previous iteration, then go to Step 6. Otherwise, go to Step 4.
4. Test all items one by one, omitting the items from the set obtained in Step 2, except when the DIF item in question is being tested. For IRT-based methods, DIF items are discarded during the rescaling of the item parameters to a common metric. For non-IRT-based methods, the DIF items are discarded from the calculation of the total test scores and related DIF measures. That is, items that are identified as DIF are omitted, and the scale or total score is recalculated.
5. Define a set of DIF items on the basis of the results of Step 4 and go to Step 3.
6. Stop.

Furthermore, methods for detecting item bias have proliferated in recent years and have been reviewed by Petersen (1977), and Rudner (1977). The various methods include techniques that examine (a) differences in relative item difficulty across different groups, (b) differences in item discrimination across groups, (c) differences in the item-characteristic curves for different groups, (d) differences in the distribution of incorrect responses for various
groups, (e) differences in multivariate factor structures across groups (Subkoviak et al, 1987).

According to Wainer, Bradlow, and Wang (2010), there are several important criteria that can be used to judge any proposed detection scheme:

1. The accuracy: it should identify items with DIF accurately and, conversely, not misidentify items that do not have DIF.
2. The efficiency: it should use the available information well and so perform its detection with as small a sample size as possible.
3. Practicality: it should work cheaply, for if it is too expensive its likelihood of being used broadly is diminished.
4. Flexibility: it should be able to detect differences in characteristics in situations that may not have been foreseen. It should also work when the assumptions underlying it are violated, even massively.

In this book, we classified DIF methods into two categories: those relying on an IRT logistic model, and those not relying on IRT. We refer to these classes of methods as IRT methods and non-IRT methods, respectively. For the former, the estimation of an IRT model is required, and a statistical testing procedure is followed, based on the asymptotic properties of statistics derived from the estimation results. For the latter, the detection of DIF items is usually based on statistical methods for categorical data, with the total test score as a matching criterion. Item Response Theory (IRT) techniques provide a powerful means of testing items for bias, using what is known as differential item functioning (DIF) analysis. In contrast, Classical Test Theory (CTT) based methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups. In essence, such
methods cannot distinguish between the situation in which (a) the subgroups have different means, and the test is biased, versus (b) the means differ, but the test is not biased.

**Non IRT Based Approaches**

**Logistic Regression (LR)**

Logistic regression is helpful when you want to predict a categorical variable from a set of predictor variables (independent variables). Logistic regression also is useful when some or all of the independent variables are dichotomous; others can be continuous. Logistic regression is based on statistical modeling of the probability of responding correctly to an item by group membership and a criterion or conditioning variable.

Logistic regression (LR) relies on the following assumptions. First, binary logistic regression assumes that the dependent variable is dichotomous and, like most other statistics, that the outcomes are independent and mutually exclusive; that is, a single case can only be represented once and must be in one group or the other. Second, logistic regression requires large samples to be accurate: some say there should be a minimum of 20 cases per predictor, with a minimum of 60 total cases. Third, multicollinearity is a potential source of confusing or misleading results and needs to be assessed. Fourth, the dependent variable must be a discrete random variable. Fifth, there should be a linear relationship between the continuous variables (independent variables) and the dependent variable (outcome). Sixth, a test examinees’ answer on one item should be independent of the test examinees’ answer on any other items (local independency). Seventh, each independent variable should be measured without an error. Eighth, the errors should be uncorrelated with the
independent variable, have a mean of zero and be normally distributed. Ninth, the error variance should be constant across levels of the independent variable (homoscedasticity).

LR DIF method is a contingency table approach and has the capability of using both continuous and multiple ability estimates as well as the dichotomous ability estimates. LR DIF procedure can be used with both polytomous and dichotomous items (Agresti, 2002). Swaminathan and Rogers (1990) applied the Logistic Regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and non uniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and non uniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of non uniform DIF compared to the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). These studies also identified two major weaknesses in the LR DIF procedure: 1) the Type I error or false positive rate was higher than expected, and 2) the lack of an effect size measure.

Logistic regression has a formal mathematical equivalence to the log linear model approach of Mellenbergh (1982): Coefficients for group, total score, and interaction terms are estimated and tested for significance with a model comparison strategy. However, logistic regression is highly similar to standard ordinary least squares regression. It can be conceptualized as an equation that uses group, ability, and group-by-ability terms to predict whether
an item response is right (1) or wrong (0). This property is desirable for didactic purposes.

Logistic regression uses the examinee as the unit of analysis, and has the following form:

$$p(u/xg) = \frac{e^{(1-u)[\beta_0 - \beta_1 x - \beta_2 g - \beta_3(xg)]}}{1+e^{(1-u)[\beta_0 - \beta_1 x - \beta_2 g - \beta_3(xg)]}}$$

or

$$p(u = 1) = \frac{e^x}{1+e^x}$$

Where:

P is the probability of individuals getting an answer correct.

g: represents group membership (0 for focal group and 1 for reference group).

x: The matching group (ability: the observed total test score).

u: represents the item response value (0 for an incorrect answer and 1 for correct answer).

xg: represents the interaction between the matching variable and the group variable.

$\beta_0, \beta_1, \beta_2, \text{and } \beta_3$: correspond to the intercept and weights for the ability, group difference and interaction between group and ability, respectively.

The above equation is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership. Once the estimates of the four coefficient parameters, $\beta_0, \beta_1, \beta_2, \text{and } \beta_3$, for an item are obtained from a sample of test responses, the usual likelihood ratio chi-square tests of significance of the estimates used to examine the presence of DIF. The null
hypothesis is that \((H_0: \beta_2 = \beta_3)\). An item shows uniform DIF if \(\beta_2 \neq 0\) and \(\beta_3=0\) with 1 degree of freedom and non-uniform DIF if \(\beta_3 \neq 0\) (whether or not \(\beta_2=0\)) with 1 degree of freedom (Swaminathan & Rogers, 1990). The item favours higher ability members of the reference group and the lower ability members of the focal group if \(\beta_3 > 0\). The item favours lower ability members of the reference group and the higher ability members of the focal group if \(\beta_3 < 0\) (Jodoin & Gierl, 2001; Rogers & Swaminathan, 1993).

The logged likelihood function:

\[
\ln \left( \frac{p_i}{1-p_i} \right) = Z = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 (x \times g)
\]

Used to calculate and interpret model comparison statistics. The likelihood values range from 0 to 1 where log likelihood values range from negative infinity to zero. Reversing this range as from 0 to positive infinity by multiplying -2 provides the same interpretation with regression models.

This model strategy comparison is made by adding the ability \((x)\), group \((g)\), and interaction terms \((x \times g)\) into the model in a hierarchical order as shown in model one (baseline or reduced model): \((Z = \beta_0 + \beta_1 x)\), model two: \((Z = \beta_0 + \beta_1 x + \beta_2 g)\), and model three: \((Z = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 (x \times g))\).

Since the coefficients are estimated using maximum likelihood estimation we can test for DIF using likelihood ratio test statistics. Model 3 is the augmented (full) model and can be used to test for both uniform and non-uniform DIF simultaneously. The second model allows us to test for uniform DIF. The third (null) model is used when there is no DIF in the item. As a result of this formulation, the larger the difference between the models, the larger the improvement in the model due to the ability and group variables (Pampel, 2000). The improvement in the model can be analyzed in two ways.
First testing uniform and non-uniform DIF simultaneously, and second testing uniform and non-uniform DIF separately (alternative comparisons). In first analysis, the change between model 1 and model 3 should be tested with a chi-square statistics with two degrees of freedom. That is, one obtains the Chi-squared value for full model and subtracts from it the Chi-squared value for reduced model. The resultant Chi-squared value can then be compared to its distribution function with 2 degrees of freedom, χ², (Zumbo, 1999). The resulting two-degree of freedom Chi-squared test is a simultaneous test of uniform and non-uniform DIF (Swaminathan & Rogers, 1990). When the test statistics exceeds χ², the hypothesis of there is no DIF is rejected.

Further, to test if the item has uniform and/or non-uniform DIF we can compare the fit of the full model with the reduced model. If β₂ is significantly separated from zero, it means that the odds of answering an item are different between the two groups. If β₁ is significantly separated from zero it means that the odds of answering an item correctly increase with increased total test score. If β₃ is significantly separated from zero it means that there is non-uniform DIF (Camilli, 2006; Camilli & Shepard, 1994). With this model 1 and model 3 comparison, the interaction term may decrease the power of the LR procedure when only uniform DIF is present because one degree of freedom is lost necessarily. Although non-uniform DIF occurs with substantially lower frequency than uniform DIF (Camilli & Shepard, 1994), it is reasonable to modify the two-degrees of freedom chi-square test into separate one-degree of freedom tests. These alternative comparisons can be made between model 3 and model 2, and between model 2 and model 1. The change between models is tested separately using chi-square statistics with one degree of freedom. As such, in the first step model 3 (full model) is tested against model 2 using a
likelihood ratio test with one degree of freedom. If there is a significant
difference we have non-uniform DIF, if it is not significant we proceed to the
next step. In the second step, we test model 2 against model 1 (reduced
model). If there is a significant difference we have uniform DIF in the item,
but if the difference is not significant we conclude that the item does not
reveal DIF.

Zumbo and Thomas (1997) indicate that an examination of both the
two-degree-of-freedom Chi squared in LR and a measure of effect size is
needed to identify DIF. For an item to be classified as revealing DIF, the two-
degree-of-freedom Chi squared test in LR had to have had a p-value less than
or equal to 0.01 and the Zumbo-Thomas effect size measure had to be at least
an $R^2$ of 0.130. Zumbo and Thomas (1996) developed an index to quantify the
magnitude of DIF for the LR procedure based on partitioning a weighted least-
squares estimate of $R^2$ that yields an effect size measure. This index is
obtained, first, by computing the $R^2$ measure of fit for each term in the LR
model (i.e., test score, group membership, test score-by-group membership
interaction) and then by partitioning the $R^2$ for each of the terms. A DIF effect
size for the group membership term is produced by subtracting the $R^2$ for
Model 2 from the $R^2$ for Model 1 (i.e., $\Delta R^2 = R^2 (model2) - R^2 (model1)$).

Also, Zumbo (1999), however, suggested using a weighted least squares $R$
squared to measure the effect size, i.e. to measure the amount of uniform or
non-uniform DIF when LR is used. Therefore, the corresponding effect size
would be the $R$ squared attributable to both the group and interaction terms
simultaneously (i.e., $\Delta R^2 = R^2 (model1) - R^2 (model3)$).

There is at least three different system of categorization of DIF when
LR is used. Zumbo and Thomas (1997) proposed $\Delta R^2$ as an effect-size
measure, defined as the difference between Nagelkerke’s $R^2$ coefficients (Nagelkerke, 1991) of the two nested logistic models. For instance, the full model and reduced model are to be compared when uniform and non uniform DIF are considered simultaneously. Zumbo and Thomas proposed the following interpretation: negligible DIF if $\Delta R^2 \leq .13$, moderate DIF if $.13 < \Delta R^2 \leq .26$ and large DIF if $\Delta R^2 > .26$. Jodoin and Gierl (2001) have proposed a less conservative scale with cut-off scores of .035 and .07, instead of .13 and .26, respectively. More recently, Jodoin & Gierl (2001) have proposed using the following guidelines: negligible DIF: $\Delta R^2 < 0.035$, moderate DIF: $0.035 \leq \Delta R^2 \leq 0.070$, and large DIF: $\Delta R^2 > 0.070$

For multiple groups, any of the aforementioned methods (MH, standardization, SIBTEST, logistic regression) can be used for pair wise comparisons between each focal group and the reference group, or just between all groups. Among the non-IRT methods, the MH method has been generalized to a simultaneous test for multiple groups (Penfield, 2001; Somes, 1986), indicated as the “generalized Mantel–Haenszel” method, as suggested by Penfield (2001). The logistic regression method can also be generalized using multiple group indicators in the regression equation. This has been suggested by Millsap and Everson (1993).

Ordinal logistic regression used proportional odds ordinal logistic regression via maximum likelihood estimation within the software package R (See: Elosua & Wells, 2013).

**Mantel-Haenszel method (M-H)**

Mantel-Haenszel procedures belong to contingency tables approaches, together with logistic regression, log linear models. The Mantel-Haenszel (M-H) procedure was originally used to match subjects retrospectively on cancer
risk factors in order to study current cancer rates (Mantel & Haenszel, 1959). The procedure has since been adapted to study differential item functioning and is now the primary DIF detection device used at the Educational Testing Service (ETS; Dorans & Holland, 1992). The M-H method works by first dividing subgroups into the reference group (e.g., males) and the focal group (e.g., females). The focal group is of primary interest in the analysis and is compared to the reference group after being matched on $\theta$ (Uttaro & Millsap, 1994). The total test score usually serves as the $\theta$ estimate and the performance (i.e., item endorsement rates) of the reference and focal groups is compared at unit intervals of $\theta$ weighted by the number of examinees at each level (Scheuneman & Gerritz, 1990). From this comparison, an odds-ratio estimator can be calculated, and a $\chi^2$ test of significance can be carried out to assess the presence of DIF.

To assess the degree of DIF present, the odds-ratio estimator can be transformed onto the ETS “delta metric” (D; Dorans & Holland, 1992). The D statistic represents the difference in item difficulty for the reference and focal groups after the total score has been taken into account (Scheuneman & Gerritz, 1990). The advantage of using the D statistic to classify degree of DIF present is that the ETS has defined the values of it into a classification scheme delineated by Dorans and Holland (1992). A D value of 0.0 indicates no DIF, a positive value indicates DIF favoring the focal group (e.g., females), and a negative D value reflects DIF that favors the reference group (e.g., males). More specifically, there are three possible degrees of DIF: (a) negligible DIF, where $\chi^2$ is not significant or the absolute value of D is less than 1.0; (b) intermediate DIF, where $\chi^2$ is significant and D is between 1.0 and 1.49 in
absolute value; and (c) large DIF, where $\chi^2$ is significant and the absolute value of D is 1.5 or larger (Dorans & Holland, 1992).

The Mantel-Haenszel technique is ideal because it does not rely solely on the $\chi^2$ statistic, which can be overly sensitive when large samples are used, which is customary in DIF analyses. The D statistic not only complements the $\chi^2$ statistic, but also allows assessments of the degree of DIF to be made. The one limitation of the M-H procedure is that it may lack power to detect DIF that is not uniform across the range of $\theta$ scores (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994).

The Mantel-Haenszel method works with the item responses for the two groups (referred to in the psychometric literature as the reference group and the focal group). As described earlier, examinees are first stored into score groups according to total test score, resulting in up to $n + 1$ score groups, where $n$ is the number of items in the test. After reference- and focal-group examinees are matched on total test score, a $2 \times 2 \times S$ contingency table is formed, where $S$ is the number of different values of the total test score. At each score level, the data can be arranged as a $2 \times 2$ table. Within the $j$th score groups, a $2 \times 2$ table of frequencies, see Table 1.

### Table 1. Contingency table for an item for the reference and focal group with total test score $k$.

<table>
<thead>
<tr>
<th>Item Score</th>
<th>Correct = 1</th>
<th>Incorrect = 0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Group</td>
<td>$A_j$</td>
<td>$B_j$</td>
<td>$n_{Rj}$</td>
</tr>
<tr>
<td>Focal Group</td>
<td>$C_j$</td>
<td>$D_j$</td>
<td>$n_{Fj}$</td>
</tr>
<tr>
<td>Total</td>
<td>$m_{1j}$</td>
<td>$m_{0j}$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

$A_j, B_j, C_j,$ and $D_j$ correspond to the number of examinees in the four cells of the $2 \times 2$ table; $n_{Rj}, n_{Fj}, m_{1j},$ and $m_{0j}$ are the marginal’s. $T_j$ is the number of
examinees in the $j^{th}$ score groups who attempted the item number investigation. The Mantel-Haenszel test statistic from P.W. Holland and D.T. Thayer (1986) has the form:

$$ MH\chi^2 = \frac{\left(\sum_{j=1}^{n} (A_j - E(A_j)) \right)^2}{\sum_{j=1}^{n} VAR(A_j)} $$

Where:

$$ VAR(A_j) = \frac{n_{Rj}m_{j}m_{0j}}{T_j(T_j-1)} $$

$$ E(A_j) = \frac{n_{Rj}m_{j}}{T_j} $$

$MH\chi^2$ is distributed approximately as a chi-square statistic with one degree of freedom. The term $A_j - E(A_j)$ represents the discrepancy between the observed number of correct responses on the item by Reference group and the expected number. When the observed number is higher than the expected, $A_j > E(A_j)$ this indicates the potential for DIF in favor of the Reference group, whereas the opposite is true if $A_j < E(A_j)$. The Log Odds Ratio($\alpha_{MH}$) is a measure of association, and $\beta_{MH} = Log(\alpha_{MH})$ is a signed index. A positive value signifies DIF in favor of the Reference group, and a negative value indicates DIF in favor of the Focal group. If the null hypothesis is true, this quantity is zero.

This statistic has the chi-square distribution with one degree of freedom. Mantel-Haenszel statistics exceeding the tabulated value of the chi-square distribution at a specified level of alpha indicate that item performance in the reference and focal groups over the $(n + 1)$ score groups is consistently
different. Two aspects of special concern to potential user of the M-H technique are: (a) how many score groups to use; and (b) whether or not to include the studied item in the total raw score used to form score groups. J.D. Scheuneman (1979) recommended the use of three to six groups for her chi-square technique for assessing item bias. P.W. Holland and D.T. Thayer (1986) are recommending a two-step procedure that includes the studied item. This procedure, however, requires a preliminary DIF analysis to purify the matching criterion. Therefore, there is a need to experimental assesses how the $\alpha$ indices are affected by the inclusion and exclusion of the studied item in forming score groups.

Fedalgo and Mdeira (2008) illustrate a unified framework for the analysis of differential item functioning using the Mantel-Haenszel methods. This is done by means of the generalized Mantel-Haenszel statistic for the analysis of the general case of $Q$ contingency tables with dimensions $R \times C$. Moreover, with the new formulation in consideration, they suggest new applications and research lines in relation to the statistics proposed. The first and natural extension of this methodology will be therefore the analysis of DIF in multiple groups ($R > 2$), that is, the application of the statistics presented to the general case of tables of dimensions $R \times C$.

**Root-mean weighted squared differences (Standardized p-difference)**

Root-mean weighted squared differences procedure is a non-parametric contingency table method used for measuring the effect size of DIF (Dorans & Kulick, 1986). The idea is to combine the difference in proportion of test examinees who answer an item correctly across the focal and reference group given their levels of total test scores. In this method, we use a weighted average of the difference in proportions between the two groups that accounts
for the number of test examinees on each level of total test score. There are two versions; the unsigned proportion difference and the signed proportion difference indices depending on whether one takes into account the sign of the difference or not. The most commonly used is the standardized p-difference which is defined as:

\[
STD - P = \frac{\sum_{j=1}^{S} n_{Fj} \left( \frac{A_j}{n_{Rj}} \frac{C_j}{n_{Fj}} \right)}{\sum_{j=1}^{n} n_{Fj}}
\]

Where: \( A_j, B_j, C_j, \) and \( D_j \) correspond to the number of examinees in the four cells of the \( 2 \times 2 \) table; \( n_{Rj}, n_{Fj}, m_{1j}, \) and \( m_{0j} \) are the marginal (see table 1).

According this method, the item reveals DIF if the difference is either >0.10 or <-0.10 (Dorans, 1989). Also, this method uses observed test score as matching variable. The main characteristics of this approach are: it is easy to work with and give stable results (Zieky, 1993), useful for measuring the size of DIF, it is a good description and can be used for explaining the nature of DIF (Camilli & Shephard, 1994), and it is simple although it lacks a test of significance (Clauser & Mazor, 1998).

**Chi Square Type DIF Methods**

The null hypothesis is either that the proportion correct between the reference and the focal groups is the same or that their odds ratio is 1. The basic idea is to calculate a chi-square statistic on each ability level in a \( 2 \times 2 \) table and then combine them into one test statistic for all ability levels. Scheuneman’s (1979) modified chi-square DIF method. The ability dimension is divided into discrete categories with the probability of correct responses in each category assumed constant, while discrimination among items vary and the lower asymptote is typically not zero. Scheuneman (1979)
stated that “item characteristic curves for different ethnic groups can be very roughly approximated using relatively small samples…” (p. 145).

Scheuneman’s version of the chi square method is concerned not only with frequencies of persons in each category as the usual chi square is, but with the number of correct responses made by persons in each group (or subpopulation) of interest. This is evident in the degrees of freedom for this method, which is \((k - 1) (r - 1)\) where \(k\) is number of subpopulations and \(r\) is the number of score groups, or categories.

Scheuneman’s (1979) modified \(\chi^2\) formula is:

\[
\chi^2 = \sum \frac{(B_E - B_O)^2}{B_E} + \sum \frac{(W_E - W_O)^2}{W_E}
\]

Where \(B\) stands for subpopulation one \((B_E: \text{expected frequencies}, B_O: \text{observed frequencies})\) and \(W\) stands for subpopulation two \((W_E: \text{expected frequencies}, W_O: \text{observed frequencies})\). For comparison purposes the usual \(\chi^2\) formula is:

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

Where:

\(O\): is the observed frequency in a given category, and

\(E\): is the expected frequency in a given category.

When establishing ability intervals on the total score scale, several criteria need to be met. The probability of a correct response within each ability interval must be less than one, and intervals are made larger or smaller to insure that there are some incorrect responses included in each interval. Expected frequencies must be at least five and all other cells must have
somewhat large counts, a minimum of ten to twenty observed correct responses, due to small cells producing spurious results (Scheuneman, 1979).

The Chi-Square Method instead of focusing on a single-item parameter, like difficulty or discrimination, other methods compare entire distributions of responses for the two groups in question. Scheuneman's chi-square procedure is one such technique. According to her definition: an item would be considered unbiased if for persons with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual. Operationally, this definition may be restated: “an item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered”.

A chi-squared method developed by Scheuneman (1979) which was criticized by Baker (1981) for yielding values that were irrelevantly affected by the size of the sample and with no known sampling distribution. The chi-square tests have the advantage of being reliable within usual standards and being homogeneous (Ironson, 1982).

**Equal Regressions**

The most widely accepted model of test bias is the regression model. This model places bias into the context of the interpretation of test scores (that is, validity), where it should be. The model says that if different groups share the same regression line, the test is not biased. If the groups have different regression lines, then the test is biased because it is measuring different things for different groups. The model says that people with the same test scores should do equally well on some external criterion.
The Mean and Covariance Structure Model (MACS)

To evaluate measurement invariance, multiple group confirmatory factor analysis (MG-CFA) is typically performed, as well as the computation of the chi-square difference test for nested models (Elosua & Wells, 2013). The first step in the detection of DIF is to define the baseline model which is referred to as the free model. To test the invariance for item i, the factor loading associated with item i is constrained to be equal between the groups. The fit of this new model is then compared to the fit of the free baseline model by taking the difference in chi-square fit statistics. If the model with the additional constraint fits significantly worse than the free baseline model, then item i is considered to function differentially between the groups. The degrees of freedom equal the difference in degrees of freedom between the two models which equals the number of parameters being compared. For each item and for each sample replication, one chi-square difference test was computed between both the baseline model and the item-constrained model. We made one comparison for every item for each data replication, excluding the reference item (for identification purposes, the 15th item was defined as the reference item; Lubke & Muthén, 2004).

Transformed Item Difficulty

Angoff (1972) offered the delta-plot or transformed item-difficulty (TID) method. The method involves computing the difficulty or p-value (proportion of subjects getting item right) for each item separately for each group. Using tables of the standardized normal distribution the normal deviate z is obtained corresponding to the (1-p) the percentile of the distribution, i.e., z is the tabled value having proportion (1-p) of the normal distribution below it. Then to eliminate negative z-values, a delta value is calculated from the z-
value by the equation $\Delta = 4z + 13$. A large delta value indicates a difficult item. For two groups, there will be a pair of delta values for each item. These pairs of delta values can then be plotted on a graph, each item represented by a point on the graph. $\Delta$ Line can be fitted to the plot of points; and the deviation (distance) of a given point from the line is taken as measure of that item's bias; large deviations indicating much bias. The distance that each point deviates from the major axis of the ellipse must be calculated. The equation used for the major of the ellipse was $Y=AX+B$ (the best fitting line) in which: $Y$ represents focal reference delta values ($\Delta_r$), $X$ represents focal group delta values ($\Delta_f$), and:

$$B = \mu_x - A\mu_y$$

Where:
A: Represents a line slope
B: The line sector of Y-axis
$\mu_y$: The mean of delta values for focal group ($\Delta_r$)
$\mu_x$: The mean of delta values for reference group ($\Delta_f$), and

$$A = \frac{(\sigma^2_y - \sigma^2_x)^2 \pm \sqrt{(\sigma^2_y - \sigma^2_x)^2 + 4 r_{XY} \sigma^2_y \sigma^2_x}}{2 r_{XY} \sigma^2_y \sigma^2_x}$$

Where:
$\sigma_x$: The standard deviation of the deltas for focal group.
$\sigma_y$: The standard deviation of the deltas for reference group.
$r_{XY}$: The correlation between deltas for reference group and focal group.

The Perpendicular distance ($D_i$) that each point deviates from the major axis was calculated from the formula:
Where:

$D_i = \frac{A X_i - Y_i + B}{\sqrt{A^2 + I}}$

$X_i$: Represents males delta value for item i.

$Y_i$: Represents females delta value for item i.

The item with $|D_i|$ values in excess of one standard deviation reveals DIF (Osterlind, 1983). In this study, the larger $|D_i|$ is, the more biased the item. A signed transformed difficulty measure of DIF, which preserved both the direction and magnitude of DIF was obtained by attaching a positive sign to $(D_i)$ if the item reveals DIF in favor of reference group, and a negative sign if the item reveals DIF in favor of focal group.

**The Simultaneous Item Bias Test (SIBTEST)**

The Simultaneous Item Bias Test (SIBTEST) proposed by Shealy and Stout (1993). This method intended to model multidimensional data. According to this procedure, the complete latent space is viewed as multidimensional, $[\Theta = (\Theta, \eta)]$, where $\Theta$ is the target ability and $\eta$ is the extraneous ability. SIBTEST method tries to reject the following null hypothesis:

$H_0: B(T) = P_R(T) - P_F(T) = 0$, in order to support the alternative hypothesis

$H_1: B(T) = P_R(T) - P_F(T) \neq 0$

$P_R(T)$ is the probability of examinees in the reference group with true score $T$ to endorse the item; and $P_F(T)$ is the probability of examinees in the Focal group with true score $T$ to endorse the item. With the SIBTEST approach, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains the biased item and the
matching subtest contains the rest of the items. For each matching subtest score, \( k \), the corresponding subtest true score for the Reference and Focal groups is estimated using linear regression. The estimated true scores are then adjusted using a regression correction technique to ensure the estimated true score is comparable for the examinees in the Reference and Focal groups on the matching subtest. In the final step, \( B(T) \) is estimated using \( B \) which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. \( (B(T)) \) is an estimate of the amount of DIF. Roussos and Stout (1996, p. 220) suggested a range of values for interpreting \( B^* \) when the null hypothesis is rejected as follows:

- If \(|B^*| < 0.059\), DIF is considered negligible (category A).
- If \(0.059 \leq |B^*| < 0.088\), DIF is considered moderate (category B).
- If \(|B^*| \geq 0.088\), DIF is considered large (category C).

**Exploratory factor analysis Approach (EFA)**

EFA of dichotomous items at an item-level uses to evaluate construct equivalence across the two groups (reference and focal). In this method, we can use a tetrachoric correlations to extract the factors (Kubinger, 2003), and Tucker’s phi coefficient to assess the congruence of the construct(s) across the two groups. According to (Zumbo et al., 2003), Tucker’s phi coefficient is commonly used to evaluate the similarity of factors across different groups. A rule of thumb is: Tucker’s phi values higher than 0.95 are viewed as evidence of factorial similarity, whereas values less than 0.85 may indicate non-negligible incongruities (Van de Vijver & Leung, 1997). Moreover, a scatter plot will be used to assess the similarity of the factor patterns by means of
cross-plotting the factor pattern coefficients of the two groups and drawing the best line of fit.

**IRT Based Approaches**

According to Hambleton, Swaminathan, and Rogers (1991), item response theory, in general, is based on two postulates: that the performance of an examinee on a test item can be explained or predicted from a set of factors traditionally called "traits, abilities, a continuum of variation"; and that the relationship between examinees' item performance and the ability underlying performance on that item can be described as an item characteristic curve (ICC). For dichotomously scored items, the usual IRT models are the logistic models with one, two, or three parameters. We further denote them by 1PL, 2PL, and 3PL models, respectively.

IRT methods can be used to detect both uniform DIF and non uniform DIF effects. The 1PL can be used only to detect a uniform DIF, and the 2PL and 3PL are suitable for the identification of uniform and non uniform DIF. The difference of item difficulty (b-difference)

**b- Difference Index**

The simplest index available to reflect differences in item parameters is the differences in estimated b parameters for the two groups (Focal and the Reference); a positive value of the difference indicates DIF favoring the Reference group, whereas a negative value of the difference indicates DIF favoring the Focal group. The simple difference in b parameters for the two groups conveys the “size” rather than the statistical significance of the DIF (Camilli & Shepard, 1994).
To detect DIF using the one-parameter logistic model, item difficulty parameter for focal group and reference group should be extracted. And the difficulty difference defines as follow:

$$\Delta b = b_F - b_R$$

Where:

- $b_R$: Estimated difficulty parameter for Reference group
- $b_F$: Estimated difficulty parameter for Focal group
- $\Delta b$: Estimated difficulty parameter difference.

To test the significant of $\Delta b$ the statistic $d$ was defined as follow:

$$d = \frac{\Delta b}{S_{\Delta b}}$$

Where:

$$S_{\Delta b} = \sqrt{S_{RF}^2 + S_{R}^2}$$

- $S_{\Delta b}$: The standard error of $b$-difference.
- $S_{RF}^2$: The variance for estimating $b$-parameter for Focal group.
- $S_{R}^2$: The variance for estimating $b$-parameter for Reference group.

Since $d$ with normal distribution and similar to $Z$ scores, the normal probability distribution tables can be used to reference the level of significance under the null hypothesis $H_0: \Delta b = 0$ (Lord, 1980). A positive value of the difference indicates DIF favoring the Reference group, whereas a negative value of the difference indicates DIF favoring the Focal group. Furthermore, a significant value of $d$ greater than or equal 1.96 indicates DIF favoring
reference group at 0.05 level, whereas a significant value of $d$ less than or equal -1.96 indicates DIF favoring focal group at the 0.05 level (Lord, 1980).

**Area index**

Prior to the DIF analyses, the focal and reference groups have to place onto the same scale by using Mean and Sigma equating method. In this method, the scale of item parameters from focal group was placed onto the scale of item parameters from reference group by using linear transformation. After placing item parameters of focal group onto the scale of item parameters of reference group, the probability of correct response was calculated for each item in each group.

An example of the item characteristic curve approaches is the area index by Rudner (1977). It measures the area between the two ICC$_S$ of the reference and the focal groups as an index of the difference between the performances of the two groups matched on ability. The index can be computed either with the signs or without the signs for the differences. The larger the area, the larger the difference is between the two curves. Other methods involve standardizing this difference or testing the equity of the two ICC$_S$.

The area must be calculated over a specified ability interval, which is from the lower group mean minus 3 SD to the upper group mean plus 3 SD. Because there is no known sampling distribution for the area statistic under the null hypothesis of no group difference, item are typically ranked according to the values of the statistic and those with the highest values flagged as potentially biased. A cut-off value (critical area) obtained by carrying out an analysis on two randomly equivalent groups (i.e. the reference group divides randomly into two equivalent groups). Because there is no bias present, the
largest area statistic obtained serves as an indicator of the greatest value of the statistic likely to occur by chance. This approach is not ideal; however, it does provide an approximate answer to the cut-off-score determination problem (Hambleton, Swamanithan & Rogers, 1991). Visual inspection of the graphs of the ICCs will reveal whether the DIF is uniform (i.e. parallel ICCs) or nonuniform (i.e. ICCs that cross). The principal conceptual unit of IRT is the item characteristic curve. An ICC is the function that relates the probability of a correct answer on an item to the “ability” measured by the test containing the item. If the unidimensional assumption of the test is met, an item response function or item characteristic curve defined by its item parameters will remain unchanged across subpopulation groups. An ICC estimated from any group will be equal to an ICC from another, and both will be equal to the ICC estimated from responses of all examinees.

**Area Index of Two-Parameter Logistic Model**

Raju (1988) formula for the 2-parameter area index can be used to find out the area between the two curves as follow:

\[
Area = 2 \left( \frac{a_2-a_1}{Da_1a_2} \right) \ln \left( 1 + e^{D(a_2-a_1)\left(\frac{b_2-b_1}{a_2-a_1}\right)} \right) - (b_2-b_1)
\]

Where:

- \(a_1\): discrimination parameter for Reference group
- \(a_2\): discrimination parameter for Focal group
- \(b_1\): difficulty parameter Reference group.
- \(b_2\): difficulty parameter for Focal group
- \(D\) = 1.7 (constant: scaling factor).
Area index is powerful in detecting nonuniform DIF. Area-based statistics rest on the premise that when an item is not reveal DIF, the ICCs for two subgroups are identical, and the area between the curves is zero. However, when an item reveals DIF, the ICCs are not the same, the area between the curves is not zero, and DIF is present (Hambleton, Swaminathan, & Rogers, 1991). The most important aspect of area index is also the most difficult to attain.

In order to accurately calculate the area between two item characteristic curves, both curves must be on the same metric, otherwise, observed large areas may be due to scaling differences rather than actual DIF. This problem is referred to as the “linking” problem (Harvey & Greenberg, 1996), and it arises whenever item parameters are estimated using data from two different subgroups (samples) of examinees (Stocking & Lord, 1983). Item DIF studies using area index method will always require subgroup parameters to be linked.

**Area Index of Three-Parameter Logistic Model**

Raju (1988) formula for the 3-parameter area index can be used to find out the area between the two curves as follow:

\[
Area = \left(1-c\right) \frac{1}{2} \left(\frac{a_2-a_1}{Da_1} \right) \ln \left[ 1 + e^{\frac{D(a_2-a_1)}{a_2-a_1}} \right] - \left(\frac{b_2-b_1}{a_2-a_1}\right) \]

Where:

- C: guessing parameter (the mean of item guessing parameters for the two groups)
- a1: discrimination parameter for Reference group
- a2: discrimination parameter for Focal group
b₁: difficulty parameter Reference group.
b₂: difficulty parameter for Focal group
D=1.7 (constant: scaling factor).

Since the c parameter for the two groups are not the same, the significance test for the area statistic cannot be carried out. The problem is to find "cut-off" values for the area statistic that can be used to decide whether DIF is present. An empirical approach to determining a cut-off is to divide the group with the larger sample size into two randomly equivalent groups, to estimate the ICCₜ in each group separately, and to determine the area between the estimated ICCₜ (Hambleton & Rogers, 1989). Since the groups are randomly equivalent, the area should be zero. Nonzero values of the area statistic are regarded as resulting from sampling fluctuations, and the largest area value obtained may be regarded as the largest value that may be expected in terms of sampling fluctuation. Any area value greater than this is assumed to be "significant" and, consequently, indicative of DIF when the focal and reference groups are compared.

**Root mean squared difference (RMSD)**

Linn et al (1981) developed approach to estimate the area between the two ICCs curves of the two groups. This approach can be used with 2PL and 3PL. In this approach, the theta $\theta_j$ measures should be calibrated and equated to be in the ability range (-3, 3), then the interval (-3,3) divides into 600 equal units (the length of each unit is .01). After that, the root mean squared difference (RMSD) of the ICCs for the two group on ability continuum (-3, 3) can be extracted using the following formula:
The two ICCs for the two groups are similar if RMSD = 0 (no DIF), whereas, the two ICCs for the two groups are different if RMSD > 0 (DIF presence). To test the significance of RMSD value, the critical area can be extracted. In this case, the item reveals DIF if RMSD greater than the critical area.

**Probability difference Indices**

The differences between ICCs should be summed in the ability range of focal group. Probability difference Indices can be defined as follows:

\[ \Delta P_j = P_{rj}(\theta_j) - P_{fj}(\theta_j). \]

Where: \( \Delta P_j \) is the probability difference, \( P_{rj}(\theta_j) \) \( P_{fj}(\theta_j) \) is the probability to endorse the item \( j \) for the reference group and the focal group respectively after controlling for theta \( \theta_j \). In this approach, we can introduce two DIF measures, namely, the signed probability difference controlling for \( \theta_j \) and unsigned probability difference controlling for \( \theta_j \). The Signed probability difference controlling for \( \theta_j \) can be calculated as follows:

\[ SPD - \theta = \frac{\sum_{j=1}^{n_j} \Delta P_j}{n_F} \]
Whereas, the unsigned probability Difference controlling for \( \theta \) can be calculated as follows:

\[
UPD - \theta = \frac{\sum n_j (\Delta P_j)^2}{n_f}
\]

If the \( UPD - \theta \) is greater than SPD, ICCs will be crossing (For more detail, see: Camilli and Shepard, 1994)

**The likelihood-ratio (LR)**

The likelihood-ratio (LR) test is one of the more popular IRT procedures for detecting DIF due to its control of Type I error rate and acceptable power rates (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinber, & Wainer, 1988). In addition, the LR test can be used to detect uniform and non-uniform DIF independently. The LR test essentially compares the fit of a compact and augmented model to test for DIF between a Reference and Focal group. The compact model constrains the parameter values to be equal between the reference and focal groups (i.e., assumes no DIF is present). The augmented model allows the parameter values for one item (or a set of items) to be freely estimated in each group, constraining the remaining items to be equal between groups. DIF is assessed by comparing the overall fit of both models. If the item being tested contains DIF (i.e., the parameter values are not equal between the groups), then the overall fit for the augmented model will be much better than the overall fit for the compact model. The overall fit of the respective model is provided by -2 times log likelihood (-2 Log L) (Elosua & Wells, 2013).
Detecting DIF using 3-Parameter Logistic Model

The detection of DIF using 3-Parameter Logistic Model has been used more often than any other IRT model. Although, the parameters in other IRT models can be estimated accurately, the assumptions of no guessing and constant item discrimination for all multiple-choice items in the test are very restrictive in practical situations. Some guessing and some variation in item discrimination is typically expected in practice.

Since the item parameters, item difficulty and item discrimination, estimated from different groups are initially in different scale and units, they have to be equated so that each of the item parameters in different groups have the same scale. After equating the parameters, asymptotic significance tests will be used to test whether or not the ICC parameters for the two groups are the same. Consequently, a biased item (item revealed DIF) is defined as the one that has different ICCs for different groups of examinees.

As a result of DIF analysis, one type of item may be shows ability level by group interaction. This occurs when the item discrimination parameter estimated from two different groups are different. Thus, ICCs will not be parallel, and may even cross. Figure 1 illustrates a clear case in which the two ICCs, one for focal group (females) and one for reference group (males), cross at the ability of -.6.

Rasch Model Approach

As we know, under the Rasch model, the guessing parameters are all assumed to be zero and items discrimination are constant (i.e. equal 1). The person ability and items difficulty should be estimated for reference group and focal group using unconditional maximum likelihood or approximation method. The item responses dichotomously scored, 1 if the response correct
and 0 otherwise. For the item calibration, it is desirable to exclude the outliers (extremes).

In this approach, the item difficulty estimates are compared across the comparisons groups (focal and reference) using asymptotic significance test distributed as a chi-square. The parameter estimates, squared standard error of estimates, and test statistics have to compute using the programs (e.g., BICAL) and FORTRAN). Further, the item characteristic curves ICCs of the reference and focal group are visually inspected. Each curve, together the overall test statistics, for each significant item at .001 level investigates by computing a pairwise comparison. Finally, we have to construct a 99.9% confidence interval on the item difficulty difference between the focal and reference groups for each item.

To test the significant of DIF using the Rach model, item difficulty parameter for focal group and reference group should be extracted. And the difficulty difference defines as follow:

\[ \Delta b = b_F - b_R \]

Where:

- \( b_R \): Estimated difficulty parameter for Reference group
- \( b_F \): Estimated difficulty parameter for Focal group
- \( \Delta b \): Estimated difficulty parameter difference.

To test the significant of \( \Delta b \) the statistic \( d \) was defined as follow:

\[ d = \frac{\Delta b}{S_{\Delta b}} \]

Where:
\[ S_{\Delta b} = \sqrt{S_{b}^2 + S_{r}^2} \]

\( S_{\Delta b} \): The standard error of b-difference.
\( S_{b}^2 \): The variance for estimating b-parameter for Focal group.
\( S_{r}^2 \): The variance for estimating b-parameter for Reference group.

Likewise a b-difference approach, a positive value of the difference indicates DIF favoring the Reference group, whereas a negative value of the difference indicates DIF favoring the Focal group. Furthermore, a significant value of \( d \) greater than or equal 1.96 indicates DIF favoring reference group at 0.05 level, whereas a significant value of \( d \) less than or equal -1.96 indicates DIF favoring focal group at the 0.05 level (Lord, 1980).

A new method for detecting differential item functioning in the Rasch Model proposed by (Strobl, Kopf & Zeileis, 2010), this approach provides no straightforward interpretation of the groups with respect to person characteristics. They propose a new method for DIF detection based on model-based recursive partitioning that can be considered as a compromise between those two extremes. By this approach we will be able to detect groups of subjects exhibiting DIF, which are not pre specified, but result from combinations of observed covariates (for more detail, refer to this report).

**Partial Credit Approach (PC)**

PC designed to use with responses that are scored in ordered categories, that is, for item responses that can be awarded partially credit. According to PC, items calibrated have associated with them several difficulty values, each indicating the difficulty of reaching a “step” along the way to successful completion of an item. That is to say, in this method, we compared between
the two groups for each step. Further, ICCs, information function, ability estimates, fit statistics, and standard errors at each ability can be extracted and used for item (steps) and test evaluation. To detect DIF of each step, Draba’s statistics (Crocker & Algina, 1986, p. 130) can be calculated using the following formula:

\[ B = \frac{d_R - d_F}{\sqrt{(SEd_R^2 + SEd_F^2)}} \]

Where:

- \(d_R\) is a step difficulty value for reference group.
- \(d_F\) is a step difficulty value for the same step in a focal group.
- \(SEd_R\): is a calibrated standard errors of step difficulty for reference group.
- \(SEd_F\): is a calibrated standard errors of step difficulty for focal group.

Since \(D\) with normal distribution and similar to \(Z\) scores, the normal probability distribution tables can be used to reference the level of significance under the null hypothesis \(H_0: \Delta D = 0\). A positive value of the difference indicates DIF favoring the Reference group, whereas a negative value of the difference indicates DIF favoring the Focal group. Furthermore, a significant value of \(D\) greater than or equal 1.96 indicates DIF favoring reference group at 0.05 level, whereas a significant value of \(D\) less than or equal - 1.96 indicates DIF favoring focal group at the 0.05 level (Lord, 1980).

References


Chapter Three
Detecting DIF using Logistic Regression and Mantel-Haenszel Approaches

Abstract

The study was conducted to find out the agreement between two approaches (i.e. logistic Regression model, and M-H) in detecting a gender-related differential item functioning of a mathematical ability scale items. The scale was developed and administered to samples of 800 males and females students (380 males and 420 females) in Jordan. The study pointed out: (1) the percentage of agreement between the two approaches in detecting DIF was 80%. (2) Males outperformed females in spatial and deductive abilities, whereas females outperformed males in numerical ability.

Introduction

That a test not be biased is an important consideration in psychological testing. That is, it is essential that a test is fair to all examinees, and is not biased against one group of examinees. Bias can result in systematic errors that distort the inferences made in test results. In many cases, test items are biased due to the fact that they contain sources of difficulty that are irrelevant to the construct being measured, and these irrelevant factors affect performance. Perhaps the item is tapping a secondary factor or factors over-and above the one of interest. This issue, known as test bias, has been the subject of a great deal of recent research, and a technique called Differential Item Functioning (DIF) analysis has become the new standard in psychometric bias analysis.
In recent years, educators have been redefining the goals of instruction and learning to include increased attention to high-level thinking skill (e.g. National Council of Teaching in Mathematics, 1989). At the same time, educators and psychometricians have been reevaluating how best to assess students’ thinking and reasoning skills. Consequently, there has been an increased interest in the use of performance assessments because they have the potential for allowing students to display their solution processes and reasoning. However, evidence is needed to ensure reliable and valid assessments of students’ high-level thinking skills. In particular, evidence is needed to ensure that inferences made from performance assessments are equally valid for different subgroups in the population; therefore, the detection of Differential Item Functioning (DIF) is important in addressing issues regarding the quality of the assessments instrument.

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1990). When personal attributes, such as gender systematically affect examinee achievement on an item, the result can be differential item functioning (DIF).

Simulation studies from educational testing experts have found that LR-based DIF detection techniques enables the detection of both uniform and non-uniform DIF, while Mantel-Haenszel techniques are better suited for the analysis of uniform DIF (Sells, 1973; Anogoff & Ford, 1973).

Several approaches have been promulgated for the statistical assessment of DIF. Most approaches or DIF were developed in educational settings in which items are generally dichotomously scored as correct or
incorrect. Mantel–Haenszel (MH) - based techniques were initially applied to the problem of assessing DIF. It was recognized by the early 1990s that logistic Regression (LR) based techniques were more powerful than MH-based techniques (Sells, 1973; Anogoff & Ford, 1973).

**Gender differences in mathematical ability**

One line of research focused on the relationship between three cognitive abilities (verbal, quantitative, and visual-spatial abilities) and gender differences in mathematical ability. However, evidence from these studies is inconsistent and sometimes conflicting. Spatial abilities generally refer to skill in representing, transforming, generating and recalling symbolic, nonlinguistic information (Linn & Petersen, 1985). Spatial skills involve the ability to think and reason using mental pictures rather than words (Nuttall, Casey, & Pezaris, 2005). They are believed as one important component of mathematical thought during mathematical problem solving (Battista, 1990; Casey, 2003; Halpern, 2000).

Spatial visualization has been defined as those spatial tasks which involve complicated multistep manipulations of spatially presented information (Linn & Petersen, 1985). Although many researchers have found that spatial visualization and problem solving were related (Battista, 1990; Fennema & Tartre, 1985; Sherman, 1979). Studies investigating gender differences in spatial visualization have reported inconsistent results. For instance, Ben- Chaim et al. (1988) found that there were statistically significant gender differences in spatial visualization among middle school students; while other researchers concluded that gender differences in spatial visualization were small or null among middle school students (Armstrong,
Mental rotation refers to the ability to transform mentally and manipulate images when the object is rotated in three dimensional space (Nuttall et al., 2005). Many studies suggested that there was a large gender difference in mental rotation ability with males outperforming females (Casey et al., 1995; Halpern, 2000; Linn & Petersen, 1985; Masters & Sanders, 1993; Voyer, et al., 1995). Evidence from a variety of sources has shown that there were gender differences in verbal skills with females outperforming males on many verbal tasks (Maccoby & Jacklin, 1974; Halpern, 2000). However, Hyde and Linn (1988) concluded that gender differences in verbal abilities had declined and were negligible now.

Studies that reported gender differences in mathematical abilities favoring males had generally consistent conclusions. Linn and Hyde (1989) concluded that females are superior at computation at all ages and that differences favoring males on problem solving emerge in high school. Benbow and Stanley (1980) indicated that gender difference in mathematical reasoning ability in favor of boys was observed before girls and boys started to differ in mathematics courses taking. This gender difference even increased through the high school years. Benbow and Stanley (1983) also suggested that males dominated the highest end on mathematical reasoning ability before they entered adolescence.

Some studies have attributed gender differences in quantitative SAT achievement to males and females’ differential patterns of course taking. They suggested that increasing female’s high-level mathematical course-taking would effectively increase their achievement in quantitative SAT.
Students taking higher level mathematics courses would benefit from training in abstract reasoning, from computational practice, and from generally being more comfortable in working with numbers. (Pallas & Alexander, 1983). This explanation was in conflict with the conclusion of Benbow and Stanley (1980), who found that gender difference in mathematical reasoning ability in favor of boys, was observed among gifted youth before they started to differ in mathematics courses taking. The inconsistent conclusion might be due to the different samples they used.

Methods for detecting item bias have proliferated in recent years and have been reviewed by Petersen (1977), and Rudner (1977). The various methods include techniques that examine (a) differences in relative item difficulty across different groups, (b) differences in item discrimination across groups, (c) differences in the item-characteristic curves for different groups, (d) differences in the distribution of incorrect responses for various groups, (e) differences in multivariate factor structures across groups (Subkoviak et al, 1987). Thus, the researcher wishing to select a bias detection method is confronted with many methods and no clear guidelines for choosing among them. The comparison of bias methods is an important practical concern. Rudner (1977) and Scheuneman (1979) have noted the need to empirically compare the various methods.

**Significance of the study**

The significance of this study is related to the importance of mathematical ability in the current mathematics education reform and the goal of achieving equal educational outcomes in students' learning of mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 2000). Since
mathematics is no longer just a prerequisite subject for prospective scientists and engineers but is a fundamental aspect of literacy for the twenty-first century (Mathematics Sciences Education Board, 1993; NCTM, 1989), male and female students should have equal opportunity to learn mathematics, have equal treatment with in classrooms, and achieve equal mathematics educational outcomes (Fennema & Leder, 1990).

The uniqueness of this study was its investigation of gender differences of relatively large samples of Jordanian students using mathematical ability test.

Research questions

The present study sought answers to the following questions: (1) to what extent do the two methods (i.e. logistic regression model and Mantel-Haenszel) agree or disagree in the identification DIF? (2) Are there gender differences in mathematical ability? (3) Are gender differences linked to content areas within mathematics?

Participants

A total of 800 (380 males and 420 females) tenth grade students in Jordan were targeted as participants in this study, during the ending period of the first semester, school year 2009-2010.

Instrument

A mathematical ability scale was developed as a part of this study. The scale compressed of 30 multiple-choice items to measure three components of mathematical ability (i.e. numerical ability, deductive ability, and spatial ability). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be
Spearman-Brown Correction on split-half reliability for odd even comparison also show similar results \( r = 0.89 \). Validity of the instrument was shown using inter-correlation of the scale (.19 to .855). Factor Analysis reveals that the test measure one trait (unidimensionality).

**Detecting DIF**

In the present study, two techniques have been used (i.e. logistic regression, and Mantel-Haenszel).

**Results and Discussions**

Logistic Regression method was used to identify Differential Item Functioning on the mathematical ability scale for each of the thirty items. According to the criteria of this approach, eighteen of thirty items or 57 percent of the items revealed DIF (ten items revealed uniform DIF, whereas eight items revealed non uniform DIF). Further, ten items were in favor of males, whereas, eight items were in favor of females.

Mantel-Haenszel method was used to identify Differential Item Functioning on the Mathematics Ability Scale for each of the thirty items. According to the criteria of this approach, the M-H procedure flagged sixteen of thirty or 53 percent of the thirty items as indicating DIF (six items were in favor of females and ten items were in favor of males).

In order to inspect the consistency between the two approaches in detecting DIF for the test, the percentage of agreements between the two approaches were computed (i.e. the degree of correspondence between the two approaches with respect to the items revealing or not revealing the DIF for all items were computed). The two methods were agreeable in allocating fourteen items as revealing DIF, and ten items as not revealing DIF. As such, the
percentage of agreement between M-H and Logistic regression methods is 80% (i.e. 14+10/30=80%) (Table 1).

**Table 1**

*Pair wise agreement between M-H and logistic regression methods.*

<table>
<thead>
<tr>
<th>Results From Logistic regression</th>
<th>No. of nonflagged</th>
<th>No. of flagged</th>
<th>Marginal total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of nonflagged items</td>
<td>10</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>4</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Marginal total</td>
<td>14</td>
<td>16</td>
<td>30</td>
</tr>
</tbody>
</table>

In summary, the percentage of agreement between the two approaches in detecting DIF are relatively high, however, this may due to: both methods related to classical test theory of measurement. This finding seems to be consistent with the previous studies (e.g.: Intaswan, 1979; Seong & Subkoviak, 1987; Hambleton & Rogers, 1989; Baghi & Ferrara, 1989; Skaggs & Lists, 1992; Hakim & Cohen, 1995; Stage, 2000).

The DIF indices point to the conclusion that females had an advantage over males on the numerical ability, whereas males had an advantage on items involving spatial ability and deductive ability. The tendency for males to perform better than females on spatial ability and inductive ability, and women to perform better on numerical ability is consistent with previous findings (e.g. Willson, Fernandez & Hadaway, 1993; Gallagher et al, 2000).

In previous studies, however, females usually performed better on Number and Computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance The Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, female show superiority in computation based on the data available. Gender differences in achievement in
mathematics in favor of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

There are many studies that focus on differences between men and women in tests (Gallagheet al, 2000; Kimball, 1994; Willingham & Cole, 1997).

From the findings of the present study and the earlier studies, one conclusion can be drawn is that males have a better spatial ability than females (Geary, 1996). Males use this spatial more often than females when solving problems, which can give advantages while solving certain kinds of problems in geometry (Geary, 1996). Many studies indicated that women are better than men in verbal skills, which can give them advantages on items where communication is important. Women also score relatively higher on tests in mathematics that better match coursework. Men tend to outperform women in geometry, arithmetic, and algebraic reasoning questions. Women tend to be better at intermediate algebra and arithmetic and algebraic operations (Willingham & Cole, 1997). Gallagher et al (2000) found men outperformed women in all kind of problems, but that the differences were greater for problems requiring spatial skills or multiple solution paths than for problems requiring verbal skills or containing classroom-based content.

Spatial abilities were reported to have relationship with mathematics test scores (Caseyet al, 1995; Geary, Saults, Liu, & Hoard, 2000; Nuttall et al, 2005). This relationship indicates that gender differences in spatial abilities may contribute to gender differences in mathematical problem solving.
This study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum.

References


Chapter Four

Detecting DIF: Comparisons between CTT and IRT Methods

Abstract

Assessment of test bias is important to establish the construct validity of tests. Assessment of differential item functioning (DIF) is an important first step in this process. DIF is present when examinees from different groups have differing probabilities of success on an item, after controlling for overall ability level. The study was conducted to answer the following questions: To what extent do the four methods (i.e. area differential index procedure for the 2- parameter Logistic model, TID, b-difference and Chi-square) agree or disagree in the identification of DIF? Are there gender differences in mathematical proficiency? What is the content or nature of those items identified as revealing DIF? Achievement test covering the following subjects: Relations and functions, polynomial, Trigonometric functions, and triangles was developed. The test was administered to a sample of 1228 tenth grade students (656 males and 624 females) in Jordan. The study pointed out: (1) the percentage of agreement among the four methods in detecting DIF was from 41% to 85%. The highest agreement was between Chi-square and b-parameter difference methods (85%), whereas the lowest agreement was between Area index and TID methods (41%). The agreement among IRT based methods and CTT based methods was convergent. (3) Females showed a statistically significant and consistent advantage over males on items involving Relations and functions, polynomial, Trigonometric functions, whereas men showed a less consistent advantage on items involving triangles, however it was
concluded that gender differences in mathematics may well be linked to content.

**Introduction**

It is important that the tests must be free of systematic demographic subgroup bias. Item response theory (IRT) techniques provide a powerful means of testing items for bias, using what is known as differential item functioning (DIF), as well as assessing the cumulative effect of any item-level bias on the test’s total score.

In contrast, classical test theory (CTT) methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups, or on differential item-passing/endorsement rates between subgroups. In essence, such methods cannot distinguish between the situation in which (a) the subgroups have different means, and the test is biased, versus (b) the means differ, but the test is not biased (i.e., one group truly has a higher average on the test).

Gender differences in mathematics have been a popular but complex issue in educational research (Fenema & Leder, 1990; Leder, 1990). Since Sells (1993) expressed the concern that mathematics is the critical filter for the differential representation of women and men in scientific and technical fields, there has been increased interest in research about gender and mathematics. In particular, researchers have focused on investigations of gender-related performance differences in mathematics and have provided different theoretical models to explain the gender differences in mathematics from various perspectives, such as biological, educational, and sociological.
Although recent reviews of research on gender differences in mathematics by Friedman (1989) and Hyde et al (1990) suggested that gender differences in mathematical performance are declining, female students continue to show less confidence in their mathematical ability and a lower perception of the usefulness of mathematics to them in the future. Even among the mathematically gifted students, females have lower educational aspirations in mathematics and sciences than do males (Benbow, 1992).

In the past, researchers have explored how the gender differences in mathematics were related to various levels of tasks and age groups. Researchers consistently found that male students are superior in geometry and visualization. On the other hand, female students show superiority in computation based on the data available.

With respect to the gender differences in mathematical problem solving, however, there are mixed results. For example, Marshall (1984) examined general differences of sixth grade students' mathematical performance in solving computation (involving whole numbers, fractions, and decimals) and word problems. She found that female students are more likely than male student’s to perform computations successfully, while male students are more likely than female students to solve word problems successfully. In another study, Marshall and Smith (1987) explored the gender differences of third grade and sixth grade students on various tasks, including computation problems, word problems, and nontraditional problems. According to Marshall and Smith (1987), third grade female students performed better than male students for both computation tasks and nontraditional problems, but there is no significant gender difference on word problems. Sixth grade female students again performed better than male students for computation tasks, but
there were not significant differences on word problems and nontraditional problems.

**Research questions**

The present study sought answers to the following questions: The first question of interest was: To what extent do the four methods (i.e. area transformed item difficulty, b-parameter difference, and Chi-square) agree or disagree in the identification DIF? A second question was: Are there gender differences in mathematical proficiency? A third question was: Are gender differences linked to content areas within mathematics

**Description of the Test Data and Examinees Samples**

A mathematical proficiency test was developed in order to measure four components of the mathematical proficiency: Relations and functions, Polynomial, Trigonometric functions, and Triangles. The primary form of the scale (60 items) was tried out to a sample of 144 students-males and females, chosen from tenth grade to make sure that the items of the test are clear and are understood by those who were tested, and to recognize the levels of difficulty and discrimination and the effectiveness of the detractors of the items. Accordingly, the final version of the scale compressed of 54 items. The test of the mathematical proficiency was applied during the last quarter of the school – year 2009/2010 to sample of (1228) students- males and females from the tenth grade (656 males, and 624 females).

The item analysis revealed levels of difficulty from 1.6 to 0.96 and levels of discriminate ability from 0.19 to 0.56. Besides, it revealed that the detractors were reversal to the item discriminate.
Data about validity of the test were collected through four methods: Internal consistency, item analysis, Logical judgment, and Factor analysis. Cronbach alpha method was used to collect data about the reliability of the test (alpha =.91). Confirmatory Factor Analysis reveals that the data obtain fits the model, and the test measures a single trait (unidimensionality).

**DIF Detection Procedures**

Four methods were used to investigate DIF (area index of the two-parameter logistic model, transformed item difficulty, b-parameter difference, and Chi-square).

**Results and Discussion**

The TID method was used to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty-four items. According to the criteria of this method, nineteen or 35 percent of items revealed DIF (two items were in favor of males and seventeen items were in favor of females). The range of D signifies DIF in favor of males were from -1.03 to -1.02, whereas the significant value of D for female students was from 1.01 to 1.91. Item difficulty (p) for each item indicates that the test is easier for females.

The b-parameter difference method was used to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty-four items at the .05 level of significance. According to the criteria of this method, forty-one or 75 percent of items were easier for females (i.e. the lowest value of b-parameter for one group indicates that the item is easy for this group), as such, the test is easier for females.

The range of b-parameter difference signifies DIF in favor of males were from .189 to 1.708, whereas the range of b-parameter difference signifies
DIF in favor of females were from -.717 to -.163. Thirty-two or 56 percent of fifty-four items revealed DIF (six items were in favor of males and twenty six items were in favor of females).

The area index was used to identify Differential Item Functioning on the mathematics proficiency test for each of fifty-four items. Forty-four or 77 percent of items revealed DIF (i.e. the area between the two curves were greater than a critical value; the critical value was .222). In order to inspect the direction of DIF (i.e. uniform or non uniform), item characteristic curves of each item for males and female were visually inspected. The item characteristic curves show that twenty nine items revealed uniform DIF (twenty four items were in favor of females and five items were in favor of males), whereas thirteen items revealed non uniform DIF. The area between the two curves for three items were closed to zero (See figures 1, 2, and 3).

Figure 1: Item reveal uniform DIF
The Chi-square method was used to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty four items at the .05 level of significance. Twenty-six or 50 percent of items
revealed DIF (five items were in favor of males and twenty one items were in favor of females).

In order to inspect the consistency between any two of the four methods in detecting DIF, the percentage of pair wise agreements among the four approaches were computed (i.e. the degree of correspondence among each of pair wise methods with respect to the items revealing or not revealing DIF for all items were computed).

The Area index and Chi-square methods were agreeable in allocating twenty-three items as revealing DIF, and seven items as not revealing DIF. As such, the percentage of agreement between Area index and Chi-square methods is 56% (i.e. $\frac{7 + 23}{54}=56\%$) (See table 1).

Table 1
Pair wise agreement between Chi-square and Area index methods.

<table>
<thead>
<tr>
<th>Results From Area index</th>
<th>No. of Non flagged</th>
<th>No. of flagged</th>
<th>Marginal Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td>7</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>3</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>Marginal total</td>
<td>10</td>
<td>44</td>
<td>54</td>
</tr>
</tbody>
</table>

The b-difference and Chi-square methods were agreeable in allocating twenty-five items as revealing DIF, and twenty-one items as not revealing DIF. As such, the percentage of agreement between b-difference and Chi-square methods is 85% (i.e. $\frac{21 + 25}{54}=85\%$) (See table 2).
Table 2
Pair wise agreement between Chi-square and b-difference methods.

<table>
<thead>
<tr>
<th>Results From Chi-square</th>
<th>Results From b-difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Non flagged</td>
</tr>
<tr>
<td>No. of non flagged items</td>
<td>21</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>7</td>
</tr>
<tr>
<td>Marginal total</td>
<td>28</td>
</tr>
</tbody>
</table>

The TID and Chi-square methods were agreeable in allocating sixteen items as revealing DIF, and twenty-three items as not revealing DIF. As such, the percentage of agreement between TID and Chi-square methods is 72% (i.e. $\frac{23 + 16}{54} = 56\%$) (See table 3).

Table 3
Pair wise agreement between Chi-square and TID methods.

<table>
<thead>
<tr>
<th>Results From Chi-square</th>
<th>Results From TID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Non flagged</td>
</tr>
<tr>
<td>No. of non flagged items</td>
<td>23</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>4</td>
</tr>
<tr>
<td>Marginal total</td>
<td>27</td>
</tr>
</tbody>
</table>

The Area index and b-difference methods were agreeable in allocating twenty-seven items as revealing DIF, and five items as not revealing DIF. As such, the percentage of agreement between Area index and b-difference methods is 59% (i.e. $\frac{5 + 27}{54} = 59\%$) (See table 4).

Table 4
Pair wise agreement between b-difference and Area index methods.

<table>
<thead>
<tr>
<th>Results From b difference</th>
<th>Results From Area index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Non flagged</td>
</tr>
<tr>
<td>No. of non flagged items</td>
<td>5</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>5</td>
</tr>
<tr>
<td>Marginal total</td>
<td>10</td>
</tr>
</tbody>
</table>
The Area index and TID methods were agreeable in allocating sixteen items as revealing DIF, and six items as not revealing DIF. As such, the percentage of agreement between Area index and TID methods is 41% (i.e. $\frac{6 + 16}{54} = 41\%$) (See table 5).

**Table 5**

Pair wise agreement between TID and Area index methods.

<table>
<thead>
<tr>
<th>Results From Area index</th>
<th>Results From TID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Non flagged</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of non flagged items</td>
<td>6</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>4</td>
</tr>
<tr>
<td>Marginal total</td>
<td>10</td>
</tr>
</tbody>
</table>

The TID and b-difference methods were agreeable in allocating seventeen items as revealing DIF, and twenty items as not revealing DIF. As such, the percentage of agreement between TID and b-difference methods is 69% (i.e. $\frac{20 + 17}{54} = 69\%$) (Table 6).

**Table 6**

Pair wise agreement between TID and b-difference methods.

<table>
<thead>
<tr>
<th>Results From b-difference</th>
<th>Results From TID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Non flagged</td>
</tr>
<tr>
<td>No. of non flagged items</td>
<td>20</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>2</td>
</tr>
<tr>
<td>Marginal total</td>
<td>22</td>
</tr>
</tbody>
</table>

**Discussion**

In summary, the percentage of agreement among the four methods in detecting DIF were from 41% to 85%. The moderate agreement was among b parameter difference and TID methods (69%). The agreement among IRT based methods and CTT based methods were convergent. The highest agreement was among Chi-square and b-parameter difference methods (85%), however, this may due to: the two methods has a statistical test of significant (standard norm to identify significant DIF). The lowest agreement was among
Area index and TID methods (41%), however, this may due to: the two methods did not have a statistical test of significant.

The theoretical reasons for the lack agreement between certain pairs of methods in the identification of DIF of items are given by Hunter (1975). He discussed several factors which may cause an item to be labeled as revealed DIF when, in fact, no DIF exists. These are (a) non-unidimensional tests, (b) differences in ability distribution of the two groups, (c) differences in item quality, (d) guessing, and (e) nonlinearity of regression. Finally, one should consider the fairness of an item in addition to its statistical index of bias. Also, this result helps to explain the low and moderate agreement reported in the measurement literature among DIF methods concerning items flagged as revealing DIF. The fact is that studies of convergence of methods for investigating DIF are influenced greatly by the unreliability of the statistics.

Hunter also claims that the chi-square approach has several problems also. The test must be unidimensional and very reliable in order for the total test score to be a valid measure of ability. In addition, this approach is very sensitive if there are differences in the total test score distributions of the groups.

Finally, Hunter notes several flaws in the item characteristic curve method. Differences in the ability distributions will be reflected in instability at the ends of the curves for different groups and different displacements of the item-characteristic curves because of unreliability. He finally states that all methods fail if the test is not unidimensional.

To better understand lack of agreement between certain pairs of methods, a closer look was taken at those items identified as revealing DIF by one method but not with another. Rudner (1976) found that all the items
identified as revealing DIF by both the transformed difficulty and b-parameter difference methods were items where the discrimination parameters were similar, but the location of the parameters differed.

In previous studies, after a statistical procedure had been used to identify potentially biased items, attempts were made to identify possible content sources of this bias. In general, this procedure has neither provided consistent nor easily generalize the results.

Chi-square method flagged the items: 9, and 23 as revealing DIF in favor of males, whereas TID method flagged the items: 25, and 26 as revealing DIF in favor of males. The two CTT methods were disagreeable in allocating the items as revealing DIF in favor of males, whereas the two methods are relatively agreeable in allocating the items as revealing DIF in favor of females. In the contrast, IRT methods were agreeable in allocating the items: 9, 52, 53, and 54 as revealing DIF in favor of males. Data analysis indicates that the four methods flagged most of items as revealing DIF in favor of females.

Mathematics items indicating DIF in favor of males were found to involve triangles and items involving real world references. Mathematics items indicating DIF in favor of females tended to involve: Relations and functions, polynomial, Trigonometric functions, miscellaneous and regular mathematics items.

For mathematics items revealing DIF in favor of male student, the content characteristics involved: solving triangles equations. The fact that this test was tied to a specific curriculum appeared to help females' performance.

Although men appear to have the advantage on mathematics achievement tests, women usually have higher average classroom grades than
men. To explain this discrepancy, Kimball (1989) speculated that men and women have different learning styles; women rely more on routine use of rules learned in class, whereas men have a more autonomous style that allows them to generalize knowledge to unfamiliar problems. Then, it might be expected that women would do better on tests that are closely linked to classroom instruction. In support of this view, Smith and Walker (1988) found that women performed slightly better than men on the ninth- and eleventh-grade New York State Regents Examination, whereas men did better on the tenth-grade paper. Although the authors explained these results by speculating that women may do better on curriculum-specific tests, it could be argued that Smith and Walker's results reflect a female advantage in algebra in the ninth grade and a male advantage in geometry in the tenth grade. Seegers and Boekaerts (1996) showed that eighth grade boys in the Netherlands performed better than girls on a mathematics test, even though the test was specifically designed to reflect classroom tests.

Such a strong female advantage in Relations and functions, polynomial, Trigonometric functions, as reflected in DIF indexes has not been previously noted. This study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum.

Furthermore, assuming that females' better performance on Relations and functions, polynomial, Trigonometric functions does indicate a reliance on algorithmic learning, females might benefit even more than males from an instructional strategy that relies less on teaching algorithms and more on teaching problem solving and effective means of approaching non-routine problems.
References


Chapter Five
DIF in Cognitive Components of Solving Mathematical Routine Problems

Abstract

DIF may be attributed to item bias but may also reflect performance differences that the test is designed to measure (Pedrajita, 2009). This study examined the gender related DIF of Malaysian students in their cognitive solution processes of solving routine mathematical problems using Maentel-Haenszel (M-H) and Logistic Regression approaches. A total of 300 sixth grade Chinese and 400 sixth grade Malay students participated in the study. The Chinese sample consisted of 144 female and 156 male students, and the Malay sample consisted of 165 female and 235 male students. A set of 31 routine items was developed. Results of the study showed that overall there were a small gender differences (favouring females) on routine problem solving for Malay and Chinese samples. Examination of students' component solutions (translation, integration, planning, and execution) of solving routine problems revealed that for the both samples there were a significant DIF on the execution component items, whereas, no significant DIF on the translation, integration, and planning components items. Furthermore, Results indicated that the percentage of agreement between the two approaches in detecting DIF is relatively high.

Introduction

When personal attributes, such as gender systematically affect examinee performance on an item, the result can be differential item functioning (DIF). Psychometricians define DIF more precisely as a situation where individuals
who have the same ability, but are members of different subgroups, do not have the same probability of a correct response to an item (Hambleton et al., 1991). Operationally, when the item characteristic curves for two or more subgroups are different, the item is showing DIF (Hambleton et al., 1991).

Gender related differential item functioning is a constant concern on large-scale standardized achievement tests in mathematics because differences between females and males are often found (e.g., Bielinski & Davison, 2001; Boughton et al., 2000; DeMars, 1998; Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997, Abedalaziz, 2010, 2011).

Presumably, because of the complexity of gender-related issues, results reported from a variety of studies are inconsistent and often even contradictory (Willingham & Cole, 1997; Hyde, 1991; Cleary, 1992). Cleary (1992) suggested that such contradictory results may be accounted for by disentangling effects of different cohorts, construct, and selectivity of the sample.

Reviews of research led to the conclusions that there were gender differences in mathematical problem solving that favored males based on the fact that male samples outperformed female samples in their studies (for example, Benbow & Stanley, 1980, 1983; Benbow, 1988; Casey et al., 1995; Gallagher & DeLisi, 1994; Royer, et al., 1999). However, these conclusions were often limited to an a typical population, normally talented or highly motivated or college bound students, and relying on the selection of measures and the particular experimental situations (Caplan & Caplan, 2005). The opposite evidence found among these high-ability populations even sometimes challenged these conclusions. For example, Pajares (1996) found that gifted girls outperformed gifted boys in mathematical problem solving.
Hyde et al. (1990) meta-analysis of 100 studies suggested that gender differences in mathematics performance were small but gender differences in mathematical problem solving with lower performance of women existed in high school and in college. Many factors such as cognitive abilities, speed of processing information; learning styles, socialization were suggested to have contributions to gender difference in mathematical problem solving (for example, Duff, Gunther, & Walters 1997; Kimball 1989; Linn & Petersen, 1985; Maccoby & Jacklin, 1974; Royer, et al., 1999).

Problem situations can establish a need to know, and foster the motivation for the development of concepts (NCTM, 1989). Therefore, students should be placed into classroom problem-solving situations from the very earliest stages of mathematics learning. Thus, problem solving is a major method for mathematics knowledge acquisition rather than merely applying the new learned mathematics knowledge to solve problems. NCTM advocates that learning is led by the search to answer questions: first at an intuitive, empirical level, then by generalizing, and finally by justifying (proving).

Routine problem solving stresses the use of sets of known or prescribed procedures (algorithms) to solve problems. Gradually, students are asked to solve more complex problems that involve multiple steps and include irrelevant data. Commencing with the concrete level, students are asked to develop their own story problem situations and demonstrate the solution process with manipulative and/or pictures and later with symbols. Such problems are later presented to the class for solution One-step, two-step, or multiple-step routine problems can be easily assessed with paper and pencil tests typically focusing on the algorithm or algorithms being used.
Mayer (1987) developed a model for analyzing cognitive components in solving word problems. In his model four cognitive components involved in solving mathematical word problems were classified and analyzed: translation, integration, planning, and execution. In order to solve a problem, a student must be able to translate each statement of the problem into a mathematical sentence or an equation. This translation process requires that the student understand English sentences. Second, the student must be able to integrate each of the statements of the problem into a coherent problem representation. This integration process requires schematic knowledge. Third, the student needs to find an adequate algorithm in order to solve the problem. This solution planning requires the student's strategic knowledge. The last component requires the student to flawlessly execute the algorithm. This solution execution requires the student's procedural knowledge. This model was successfully applied to assess students' routine problem-solving skills in other studies (Mayer, Tajika, & Stanley, 1991).

In the past, researchers have explored how the gender differences in mathematics were related to various levels of tasks and age groups. Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, female students show superiority in computation based on the data available. With respect to the gender differences in mathematical problem solving, however, there are mixed results. For example, Marshall (1984) examined general differences of sixth-grade students' mathematical performance in solving computation (involving whole numbers, fractions, and decimals) and word problems. She found that female students are more likely than male students to perform computations
successfully, while male students are more likely than female students to solve word problems successfully.

Hough (2003) examined the gender differences of U.S. and Chinese students in their solution processes of solving routine and non-routine mathematical problems. Results of the study showed that overall there were statistically significant gender differences (favoring males) on both routine and non-routine problem solving for the U.S. sample, but not for the Chinese sample. However, examinations of students component processes (translation, integration, planning, and execution) for solving routine problems revealed that significant gender differences only exist for the execution component (computation skills) for the U.S. sample.

In conclusion, a large body of literature reports that there are gender differences in mathematical problem solving favoring males. The literature has consistently reported that males outperform females on mathematics problem solving among high ability students on standardized mathematics tests. These genders related differences are generally obvious in high school and in college and can be traced back to the very early stage of elementary schooling. Furthermore, these gender differences are varying across mathematical tasks. It is found that students’ strategy use is related to cognitive abilities, speed of processing information, physiological differences in brains, influences of sex hormones, learning styles, learners’ attitudes, and stereotype threat in mathematics tests, differences in socialization, and the impact of socioeconomic variables (Hembree, 1992).
Significance of the study

The significance of this study is related to the importance of problem solving in the current mathematics education reform and the goal of achieving equal educational outcomes in students' learning of mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 2000). Since mathematics is no longer just a prerequisite subject for prospective scientists and engineers but is a fundamental aspect of literacy for the twenty-first century (Mathematics Sciences Education Board, 1993; NCTM, 1989), male and female students should have equal opportunity to learn mathematics, have equal treatment within classrooms, and achieve equal mathematics educational outcomes (Fennema & Leder, 1990). The examination of gender-related performance differences on routine allows for investigating gender differences in their thinking and reasoning as they solve these problems.

Current education reform in general and mathematics education reform in particular emphasize the importance of thinking, understanding, reasoning, and problem solving in students' learning (e.g., NCTM, 1989, 1991, 2000; National Research Council, 1989). Such reform effort in mathematics curriculum and instruction requires examination of male and female students' thinking, reasoning, and problem solving rather than merely computation and symbol manipulation.

The uniqueness of this study was its investigation of gender differences of relatively large samples of Malaysian students using routine problem solving test. Moreover, the present study tries to detect a gender related DIF of cognitive processes of solving routine mathematical problems. This study provided an opportunity to examine issues in mathematics learning in general.
and issues in gender–related differential item functioning of routine problem solving in specific.

**Research questions**

The present study sought answers to the following questions: To what extent do the two methods (i.e. Mantel-Haenszel & Logistic Regression) agree or disagree in the identification DIF? Are there gender differences in cognitive components of solving routine word problems? Are gender differences linked to content areas within mathematics?

**Samples**

A total of 300 sixth grade Chinese and 400 sixth grade Malay students were participated in the study. The Chinese sample consisted of 144 (48%) female and 156 (52%) male students, and the Malay sample consisted of 165 (41%) female and 235 (59%) male students. Chinese and Malay samples are students from different public schools in Kuala Lumpur / Malaysia. Schools and students had been selected randomly during the second semester of the school year 2009-2010.

**Instrument**

Mayer's model was used to examine gender differences in the processes of solving routine problems. A set of 31 multiple-choice items was used to assess component processes of solving routine problems: five items for the translation component, five items for the integration component, 5 items for the planning component, and 16 items for the execution component. The execution component involved students' computation skills (addition, subtraction, multiplication, and division) on different types of numbers (whole numbers, decimals, and fractions).
The test was tried out on a sample of 200 students—males and females, to make sure that the items of the test are clear and are understood by those being tested, and to find out the psychometric properties of the test. Accordingly, the item analysis revealed levels of difficulty from .31 to .92. Besides, it revealed that the detractors were reversal to the item discriminate. Data about validity of the scale were collected through three methods: Internal consistency, item analysis, content validity.

The Cronbach’s alpha coefficients calculated for the Execution, Translation, Integration, and Planning subscales were .86, .70, .70, and .71, respectively, and it was calculated to be .87 for the entire scale. The scale correlation coefficients ranged between .37 and .47 on execution component, between .39 and .58 on translation component, between .39 and .46 on integration component, and between .37 and .62 on planning component. It is generally agreed that correlations in the range of .38 to .63 are useful and statistically significant beyond the 1% level, whereas correlations less than .25 are not useful and statistically non-significant (Brown, 1983). Thus, the results revealed that the alpha coefficients for all subscales were significantly high, suggesting that the internal reliability index of the four constructs and the entire scale is adequate.

**DIF Detection Procedure**

Methods for detecting DIF have proliferated in recent years and have been reviewed. The various methods include techniques that tested differences in relative item difficulty among different groups, differences in item discrimination among different groups, differences in the item-characteristic curves (ICCs) for different groups, differences in the distribution of incorrect

A plausible but not exhaustive classification of DIF detection techniques is as follows: Classical Test Theory CTT-based methods, Factor Analysis FA-based methods, $\chi^2$-based methods, and Item Response Theory IRT-based methods. According to Cole and Moss (1989), the work on DIF has focused upon the last two approaches, namely, those based upon $\chi^2$ and IRT. Unlike CTT-based methods, these last two approaches are conditional methods. In turn, $\chi^2$-based methods can be divided into four different groups: (1) the $\chi^2$-based methods in the strict sense, i.e., the $\chi^2$ correct (Scheuneman, 1979) and the $\chi^2$ full (Camilli, 1979); (2) the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), a natural outgrowth of the former $\chi^2$ methods, which is widely used and easy to implement; (3) the Loglinear Models (LM) (Mellenbergh, 1982) to test the conditional independence of group membership and the score on the studied item given the matching variable, and (4) the Logistic Regression (LR) procedure (Rogers & Swaminathan, 1993).

This paper compared the potential of two of these methods for detecting DIF (i.e. Mantel-Haenszel and Logistic Regression). Test experts and developers should use contingency table (CT) methods, particularly the LR and MH methods, in item DIF detection. These two methods are viable in the detection of DIF and are widely implemented in both test construction and research settings (Pedrajita & Talisayon, 2009).
Results and Discussion

Several items were found to contain DIF in the comparisons made. There were two comparisons: Malay males versus Malay females, and Chinese males versus Chinese females. The group means, $\Delta$, and $\chi^2$ statistic obtained in the Malay males versus Malay female’s comparison were extracted. According to the ETS criteria, six items or 19 percent of the items revealed "large DIF" (five items were in favor of females, whereas one item was in favor of males). Combining the items that exhibited either “intermediate” or “large” DIF shows that there were eight or 26 percent of the items revealed DIF (seven items were in favor of females, whereas one item was in favor of males). The value of $\Delta$ signifies DIF in favor of males was -1.54, whereas the significant range of $\Delta$ for female students were from 1.03 to 6.75.

The group means, $\Delta$, and $\chi^2$ statistic obtained in the Chinese males versus Chinese female’s comparison were extracted. According to the ETS criteria, two items or 6 percent of the items revealed "large DIF" (two items were in favor of females). Combining the items that exhibited either “intermediate” or “large” DIF shows that there were seven or 23 percent of the items revealed DIF (in favor of females). The range of $\Delta$ signifies DIF in favor of females were from 1.03 to 1.96.

The Logistic Regression method was used to identify Differential Item Functioning for each of the thirty-one items for Malay sample. Four items or 13 percent of the items revealed DIF (two items revealed uniform DIF in favor of females, whereas two items revealed non-uniform DIF). Further, the Logistic Regression method was used to identify Differential Item Functioning
for each of the thirty-one items for Chinese sample. Five items or 16 percent of the items revealed DIF (three items revealed uniform DIF in favor of females, whereas two items revealed non-uniform DIF).

In order to inspect the consistency between the two approaches in detecting DIF for Malay sample, the percentage of agreements between the two approaches was computed (i.e. the percentage of items revealing or not revealing DIF). The percentage of agreement between the Mantel-Haenszel and Logistic Regression approaches is 81% (i.e. $3+22/31 =78\%$) (See table 1).

**Table 1**
The Agreement between Logistic Regression and Mantel-Haenszel Approaches in Detecting DIF for Malay Sample

<table>
<thead>
<tr>
<th>Results From Logistic Regression</th>
<th>Results From Mantel-Haenszel approach</th>
<th>No. of Non flagged</th>
<th>No. of flagged</th>
<th>Marginal Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td></td>
<td>22</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>No. of flagged items</td>
<td></td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Marginal total</td>
<td></td>
<td>27</td>
<td>4</td>
<td>31</td>
</tr>
</tbody>
</table>

**Table 2**
The Agreement between Logistic Regression and Mantel-Haenszel Approaches in Detecting DIF for Chinese Sample

<table>
<thead>
<tr>
<th>Results From Logistic Regression</th>
<th>Results From Mantel-Haenszel approach</th>
<th>No. of Non flagged</th>
<th>No. of flagged</th>
<th>Marginal Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td></td>
<td>24</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>No. of flagged items</td>
<td></td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Marginal total</td>
<td></td>
<td>26</td>
<td>5</td>
<td>31</td>
</tr>
</tbody>
</table>

In order to inspect the consistency between the two approaches in detecting DIF for Chinese sample, the percentage of agreements between the two approaches was computed (i.e. the percentage of items revealing or not revealing DIF). The percentage of agreement between the Mantel-Haenszel and Logistic Regression approaches is 81% (i.e. $5+24/31 =93\%$) (See table 2).
This study examined the gender-related DIF of Malaysian (i.e. Malay, and Chinese) students in cognitive components of solving routine problems. The results of the differential item functioning analysis showed that there were statistically gender related DIF present in cognitive process of solving routine problems.

Results of the study showed that overall there were gender differences (favoring females) on routine problem solving for the Malay and Chinese samples. Examination of students' component solutions (translation, integration, planning, and execution) of solving routine problems revealed that for both samples there were a significant DIF on the execution component items, but not on the translation, integration, and planning components items.

The Logistic Regression and the Mantel-Haenszel Statistic yielded very similar results with respect to uniform differential item functioning (DIF). The two procedures result in similar number and identity of items being identified. Hence, there is a high degree of correspondence between these two procedures. In summary, the percentage of agreement between the two approaches in detecting DIF for both samples are relatively high (78% for Malay sample and 93% for Chinese sample), however, this may due to: both methods related to classical test theory of measurement. This finding seems to be consistent with the previous studies (e.g., Hambleton & Rogers, 1989; Skaggs & Lists, 1992; Hakim & Cohen, 1995; Abedalaziz, 2010).

The MH and LR techniques share some convenient features, i.e., ease of implementation or applicability, associated tests of statistical significance and, finally, the requirement of the number of examinees needed to get satisfactory results is not quite as large as with other conditional methods.
Further, the results of the study showed that overall there were a small
gender related DIF (favoring females) on Execution component items for
Malay and Chinese samples. The findings in the present study seem to be
consistent with the findings from previous studies. Hyde, et al., (1990)
suggested that there were very small or null gender differences in mathematics
performance on Scholastic Assessment Test-Mathematics (SAT-M) tests.
Caplan and Caplan (2005) even argued that the link between gender and the
mathematics performance was very weak.

How can we explain the finding that there were small genders related
DIF for Malaysian students in cognitive components of solving routine
problems? In Malaysian society, women and men tend to have equal
opportunities for jobs and equal salaries. Thus, the finding that there was no
gender related DIF in solving routine problems for Malaysian samples may be
explained by the fact that Malaysian students are raised in relatively more
uniform educational and social conditions. In addition, the mathematics
curriculum of sixth grade concentrates on teaching problem solving ability.

The finding that female students outperformed male students on the
execution component (measuring computation skills) in the present study
seems to be consistent with the findings from previous studies. In fact,
previous studies reported that in general, female students were more
successful than male students on computation tasks (Geary, 1996; Hyde et al.,
1990; Marshall & Smith, 1987; Abedalaziz, 2011). Only the item 20 which
measures mathematical thinking was in favor of Malay males.

Further, the finding from the present study seem to be consistent with
the findings from several meta-analyses (e.g., Wilder & Powell, 1989; Hyde et
al., 1990), which have revealed that male students usually score higher than
female students on tasks requiring mathematical thinking and problem solving.

In conclusion, the present study provides evidence that the link between gender and mathematics routine problems was very weak. Also, the study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even content are closely tied to curriculum. Furthermore, assuming that females' better performance on computation does indicate a reliance on algorithmic learning, women might benefit even more than men from an instructional strategy that relies less on teaching algorithms and more on teaching problem solving (Abedalaziz, 2010).

The conclusions drawn from this study have to be focused on the particular characteristics of the tests and samples of examinees considered. It would be useful to widen the comparison of DIF detection techniques to other conditions. Special attention should be paid to factors such as percentage of DIF items, test length, and the type and magnitude of DIF. These factors could have an important impact on DIF detection. Therefore, it would be interesting to check whether these results are generalizable to conditions other than those considered here. Among the factors which may affect the DIF results, are the number of DIF items in the test, difficulty level of items, the degree of DIF in items, the number of individuals in each group, the difference between the group means.

Since the one limitation of the M-H procedure is that it may lack power to detect DIF that is not uniform across the range of theata (θ) scores (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990), further researches are needed, to detect a gender related DIF of Malaysian students in
routine problems using different DIF methods. A natural extension of this study is to examine gender differences of low and high school students in their solution processes of solving routine mathematical problems.

The findings deserve further comment. First, the number of items exhibiting DIF with both the LR and the MH procedures seems very low. Second, consistent with earlier research, the MH and the LR procedures result in similar number of items (and similar items) being identified (Rogers & Swaminathan, 1993). Thus, there is a high degree of correspondence between the LR and the MH procedures when either one or two ability estimates are included in the analysis. LR has shown that under comparable conditions, when matching is based on a single test score, it produces results that are extremely similar to those produced using the MH Statistic. Methods for detecting DIF may be evaluated in terms of external evidence of validity. Some possible types of validity evidence for a bias technique would be a demonstration that: (1) the procedure is not selecting items at random; and (2) the results obtained with different methods tend to agree. The LR and MH procedures appear to have demonstrated the external validity evidence mentioned above. Hence, these two approaches are widely implemented in DIF detections (Pedrajita & Talisayon, 2009).

References


Chapter Six
Detecting DIF Using TID

Abstract

The purpose of the study was to examine gender differences in performance on multiple-choice mathematical ability test, administered within the context of high school graduation test that was designed to match eleventh grade curriculum. The transformed item difficulty (TID) was used to detect a gender related DIF. A random sample of 1400 eleventh graders in Kuala Lumpur was selected. In DIF indexes, females showed a statistically significant and consistent advantage over males on items involving algebra, whereas males showed a less consistent advantage on items involving geometry and measurement, number and computation, data analysis, and proportional reasoning. However it was concluded that gender differences in mathematics may well be linked to content.

Introduction

The validity of a test depends on the quality of the items included in the test or instrument. A test is never better than the sum of its items; hence to identify problematic items through item analysis is of great importance. Item analysis includes using statistical techniques to examine the test takers’ performance on the items. One important part of the item analysis is to examine Differential Item Functioning, DIF, in the items.

Gender related differential item functioning in mathematics are found; their directions seem to depend on the form of assessment used (Abedalaziz, 2010, 2011, 2012). For instance, Abedalaziz (2011), indicated that females outperformed males in algebra and arithmetic computation, whereas, males
outperformed females in spatial ability and geometry. Furthermore, Abedalaziz (2012), found out that females are often reported to outperform males on classroom grades; males are frequently outperformed females on standardised tests, timed examinations, and the most cognitively demanding mathematical tasks. The achievement measures taken as indicators of mathematical success or potential are usually those on which males excel over females. Those concerned with gender equity in mathematics learning have sought explanations for the persistent patterns of gender difference favouring males (Fennema & Leder, 1990). Research has revealed a complexity of interacting factors that include: students’ confidence and liking of mathematics, students’ beliefs about mathematics as a male domain, the reasons given for success and failure in mathematics, society, parents, the peer group, the teacher, and the classroom learning environment.

**Purpose**

The purpose of this study was to analyze gender differences in performance on multiple-choice items mathematical ability test. Because there is now a movement away from exclusive reliance on multiple-choice assessment, led by the National Council of Teachers of Mathematics (NCTM; 1994), the ability meaningfully to compare performance on different item types is important.

This study sought answer to the following question: Are there gender-related differences in mathematical performance?
Participants

A total of 1400 (690 males and 710 females) eleventh grade students in Kuala Lumpur were targeted as participants in this study, during the ending period of the First semester, school year 2009-2010.

Materials

The mathematical ability test was developed as part of this study. The 40-item instrument consists of four components (basic arithmetic, verbal arithmetic, elementary algebra, and geometry). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be .87. Spearman-Brown Correction on split-half reliabilities for odd even comparison also show similar results r= .89. Validity of the instrument was shown using inter-correlation of the sub scales (0.19 to .855).

Data Analysis

Two sections of analysis were done to establish psychometric properties. First is using the classical test theory steps which include the item analysis. Microsoft Excel was used for the analyses and computations involved in the CTT analysis. SPSS software was also used to determine reliability of the test. The second is detecting DIF.

Detecting DIF

TID approaches was used to detect DIF of mathematical test items.

Results

Table 1 shows the DIF statistic of the TID method for each of 40 items. The TID method flagged ten items at the .05 level of significance (the item 27 was in favor of female students, whereas the items 1, 14, 19, 20, 25, 33, 34, 37, and 39 were in favor of male students). Table 1 provides the p-value (item
difficulty) of test items for male and female students, which were relatively convergent.

From the analysis it appears that the effect of DIF in a well construct test was not very large. Neither group was greatly affected across all items since some items were revealed DIF in favor for each group.
Table 1
Summary Results from the TID method to Identify Differential Item Functioning on a Mathematical Ability Test

<table>
<thead>
<tr>
<th>Item</th>
<th>PM</th>
<th>PF</th>
<th>ZM</th>
<th>ZF</th>
<th>ΔM</th>
<th>ΔF</th>
<th>Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.87</td>
<td>0.83</td>
<td>-1.10</td>
<td>-0.97</td>
<td>8.60</td>
<td>9.12</td>
<td>-1.20*</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>0.62</td>
<td>-0.31</td>
<td>-0.29</td>
<td>11.76</td>
<td>11.84</td>
<td>-0.45</td>
</tr>
<tr>
<td>3</td>
<td>0.16</td>
<td>0.12</td>
<td>1.01</td>
<td>1.17</td>
<td>17.04</td>
<td>17.68</td>
<td>-0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.88</td>
<td>0.84</td>
<td>-1.20</td>
<td>1.00</td>
<td>8.20</td>
<td>17.00</td>
<td>-6.44</td>
</tr>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.89</td>
<td>-1.18</td>
<td>-1.23</td>
<td>8.28</td>
<td>8.08</td>
<td>-0.80</td>
</tr>
<tr>
<td>6</td>
<td>0.63</td>
<td>0.68</td>
<td>-0.32</td>
<td>-0.47</td>
<td>11.72</td>
<td>11.12</td>
<td>-0.56</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>0.64</td>
<td>-0.53</td>
<td>-0.36</td>
<td>10.88</td>
<td>11.56</td>
<td>-0.96</td>
</tr>
<tr>
<td>8</td>
<td>0.67</td>
<td>0.55</td>
<td>-0.45</td>
<td>-0.38</td>
<td>11.20</td>
<td>11.48</td>
<td>-0.66</td>
</tr>
<tr>
<td>9</td>
<td>0.58</td>
<td>0.64</td>
<td>-0.21</td>
<td>-0.36</td>
<td>12.16</td>
<td>11.56</td>
<td>0.03</td>
</tr>
<tr>
<td>10</td>
<td>0.38</td>
<td>0.20</td>
<td>0.86</td>
<td>0.86</td>
<td>16.44</td>
<td>16.44</td>
<td>0.30</td>
</tr>
<tr>
<td>11</td>
<td>0.23</td>
<td>0.13</td>
<td>1.12</td>
<td>1.12</td>
<td>17.48</td>
<td>17.48</td>
<td>0.45</td>
</tr>
<tr>
<td>12</td>
<td>0.64</td>
<td>0.62</td>
<td>-0.36</td>
<td>-0.31</td>
<td>11.56</td>
<td>11.76</td>
<td>-0.56</td>
</tr>
<tr>
<td>13</td>
<td>0.67</td>
<td>0.60</td>
<td>-0.43</td>
<td>-0.25</td>
<td>11.28</td>
<td>12.00</td>
<td>-0.93</td>
</tr>
<tr>
<td>14</td>
<td>0.67</td>
<td>0.50</td>
<td>-0.44</td>
<td>0.01</td>
<td>11.24</td>
<td>13.04</td>
<td>-1.61*</td>
</tr>
<tr>
<td>15</td>
<td>0.60</td>
<td>0.50</td>
<td>-0.24</td>
<td>0.00</td>
<td>12.04</td>
<td>13.00</td>
<td>-0.96</td>
</tr>
<tr>
<td>16</td>
<td>0.71</td>
<td>0.74</td>
<td>-0.56</td>
<td>-0.65</td>
<td>10.76</td>
<td>10.40</td>
<td>-0.33</td>
</tr>
<tr>
<td>17</td>
<td>0.63</td>
<td>0.70</td>
<td>-0.32</td>
<td>-0.53</td>
<td>11.72</td>
<td>10.88</td>
<td>0.11</td>
</tr>
<tr>
<td>18</td>
<td>0.51</td>
<td>0.46</td>
<td>-0.02</td>
<td>0.10</td>
<td>12.92</td>
<td>13.40</td>
<td>-0.53</td>
</tr>
<tr>
<td>19</td>
<td>0.60</td>
<td>0.55</td>
<td>-0.36</td>
<td>-0.13</td>
<td>11.56</td>
<td>12.48</td>
<td>-1.01*</td>
</tr>
<tr>
<td>20</td>
<td>0.82</td>
<td>0.71</td>
<td>-0.91</td>
<td>-0.56</td>
<td>9.36</td>
<td>10.76</td>
<td>-1.64*</td>
</tr>
<tr>
<td>21</td>
<td>0.60</td>
<td>0.53</td>
<td>-0.23</td>
<td>-0.07</td>
<td>12.08</td>
<td>12.72</td>
<td>-0.76</td>
</tr>
<tr>
<td>22</td>
<td>0.69</td>
<td>0.79</td>
<td>-0.49</td>
<td>-0.8</td>
<td>11.04</td>
<td>9.80</td>
<td>0.26</td>
</tr>
<tr>
<td>23</td>
<td>0.51</td>
<td>0.47</td>
<td>-0.02</td>
<td>0.07</td>
<td>12.92</td>
<td>13.28</td>
<td>-0.46</td>
</tr>
<tr>
<td>24</td>
<td>0.77</td>
<td>0.79</td>
<td>-0.72</td>
<td>-0.82</td>
<td>10.12</td>
<td>9.72</td>
<td>-0.40</td>
</tr>
<tr>
<td>25</td>
<td>0.61</td>
<td>0.54</td>
<td>-0.37</td>
<td>-0.11</td>
<td>11.52</td>
<td>12.56</td>
<td>-1.09*</td>
</tr>
<tr>
<td>26</td>
<td>0.64</td>
<td>0.68</td>
<td>-0.37</td>
<td>-0.48</td>
<td>11.52</td>
<td>11.08</td>
<td>-0.17</td>
</tr>
<tr>
<td>27</td>
<td>0.54</td>
<td>0.84</td>
<td>-0.10</td>
<td>-1.00</td>
<td>12.60</td>
<td>9.00</td>
<td>1.97*</td>
</tr>
<tr>
<td>28</td>
<td>0.36</td>
<td>0.42</td>
<td>0.36</td>
<td>0.19</td>
<td>14.44</td>
<td>13.76</td>
<td>0.42</td>
</tr>
<tr>
<td>29</td>
<td>0.20</td>
<td>0.18</td>
<td>0.84</td>
<td>0.90</td>
<td>16.36</td>
<td>16.6</td>
<td>0.14</td>
</tr>
<tr>
<td>30</td>
<td>0.53</td>
<td>0.54</td>
<td>-0.09</td>
<td>-0.11</td>
<td>12.64</td>
<td>12.56</td>
<td>-0.22</td>
</tr>
<tr>
<td>31</td>
<td>0.60</td>
<td>0.52</td>
<td>-0.24</td>
<td>-0.04</td>
<td>12.04</td>
<td>12.84</td>
<td>-0.86</td>
</tr>
<tr>
<td>32</td>
<td>0.50</td>
<td>0.60</td>
<td>0.00</td>
<td>-0.26</td>
<td>13.00</td>
<td>11.96</td>
<td>0.43</td>
</tr>
<tr>
<td>33</td>
<td>0.57</td>
<td>0.42</td>
<td>-0.18</td>
<td>0.20</td>
<td>12.28</td>
<td>13.80</td>
<td>-1.28*</td>
</tr>
<tr>
<td>34</td>
<td>0.67</td>
<td>0.57</td>
<td>-0.44</td>
<td>-0.19</td>
<td>11.24</td>
<td>12.24</td>
<td>-1.11*</td>
</tr>
<tr>
<td>35</td>
<td>0.26</td>
<td>0.30</td>
<td>0.66</td>
<td>0.54</td>
<td>15.64</td>
<td>15.16</td>
<td>0.48</td>
</tr>
<tr>
<td>36</td>
<td>0.29</td>
<td>0.42</td>
<td>0.55</td>
<td>0.20</td>
<td>15.20</td>
<td>13.80</td>
<td>0.99</td>
</tr>
<tr>
<td>37</td>
<td>0.46</td>
<td>0.23</td>
<td>0.10</td>
<td>0.73</td>
<td>13.40</td>
<td>15.92</td>
<td>-1.73*</td>
</tr>
<tr>
<td>38</td>
<td>0.57</td>
<td>0.48</td>
<td>-0.18</td>
<td>0.05</td>
<td>12.28</td>
<td>13.20</td>
<td>-0.90</td>
</tr>
<tr>
<td>39</td>
<td>0.34</td>
<td>0.19</td>
<td>0.41</td>
<td>0.87</td>
<td>14.64</td>
<td>16.48</td>
<td>-1.12*</td>
</tr>
<tr>
<td>40</td>
<td>0.31</td>
<td>0.31</td>
<td>0.49</td>
<td>0.50</td>
<td>14.96</td>
<td>15.00</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note. * The item reveal DIF. PM item difficulty for males. PF item difficulty for females. ΔM delta value for males group. ΔF delta value for females group. ZM z score for males. ZF z score for females.
Discussion

Both mean raw scores and DIF index point to the conclusion that females had an advantage over males on the Algebra (item number 27), whereas males had an advantage on items involving proportional reasoning, Number and Computation, Data Analysis, and Geometry and Measurement. The tendency for males to perform better than females on Geometry and Measurement and females to perform better on Algebra is consistent with previous finding (e.g., Abedalaziz, 2011, and 2012). In previous studies, however, females usually performed better on Number and Computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance (for example, Maccoby & Jacklin, 1974; Fennema & Carpenter, 1981; Halpern, 2000).

Such a strong female advantage in algebra, as reflected in DIF indexes has not been previously noted. The advantage could perhaps be explained by noting that the algebra items were very abstract and algorithmic, unlike the items in the other component, the algebra component included no real-world situation, no geometry, no word problems (Abedalaziz, 2011, 2012). As already noted, females tend to do better on such problems. Perhaps these gender related differences in performance are a result of both a reliance on routines learned in class as proposed by Bohlin (1994), Hyde and Linn (1989), Kimball (1989), and Gallagher (1992), and a lack of confidence on non routine tasks as suggested by Seegers and Boekaerts (1996).

For those five mathematics items identified by Agoff's technique as revealing DIF in favor of females, the \( P \) values ranged from .2922 to .5480 with a mean \( p \) value of .4293. The mean \( P \) value for males for all 90
The point biserial correlations for these five items tended to be in the moderate to high range, with only one item having a discrimination index below .30.

Of the eight mathematics items identified as revealing DIF in favor of males, four were verbal application items. The content of these items dealt with such concepts as determining: (a) the number of ounces of chemical to be added to each 100 gallons of water, (b) the relative altitude of two towns in a desert, (c) the number of miles per gallon of gasoline averaged by a plane, and (d) the number of gallons of orange juice remaining in a tank if a certain percentage was lost by leakage. Two other of the night mathematics items involved geometry or referred to graphs. Of the remaining two items, one item dealt with finding the sum of two times expressed in hours and minutes and the other item involved the determination of a percent.

Of the seven mathematics items identified as revealing DIF in favor of females, one item involved the multiplication of decimal fractions. The other six items required knowledge of the more abstract concepts of mathematics, such as, algebraic concepts, mathematical definitions such as, "prime numbers," "associative property," and the expression of a number in expanded notation.

References


Chapter Seven
A Gender-Related Differential Item Functioning of Mathematics Test Items

Abstract

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1997). When personal attributes, such as gender or ethnicity systematically affect examinee performance on an item, the result can be differential item functioning. The study was conducted to explore a gender-related differential item functioning in mathematics. Three methods (i.e. M-H, TID, and b-parameter difference) were used to detect DIF, and find out the agreement among these methods. The samples used in this study were drawn from a data set containing the responses of approximately 3390 (1600 males and 1790 females) eleventh grade students to achievement test comprised 45 items. In summary, the percentage of agreement among the three approaches in detecting DIF is relatively low. The range is from 43% to 65% for detecting DIF. The highest agreement was among M-H and TID methods, the lowest agreement was among TID and b-parameter difference. This study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum.

Introduction

Gender related differential item functioning is a constant concern on large-scale standardized achievement tests in mathematics because differences
between females and males are often found (e.g., Bielinski & Davison, 2001; Boughton et al., 2000; DeMars, 1998; Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997).

Presumably, because of the complexity of gender-related issues, results reported from a variety of studies are inconsistent and often even contradictory (Willingham & Cole, 1997; Hyde; 1991; Cleary, 1992) suggest that such contradictory results may be accounted for by disentangling effects of different cohorts, construct, and selectivity of the sample.

Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, female show superiority in computation based on the data available. Gender differences in achievement in mathematics in favour of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

**Research questions**

Rudner (1977) and Scheuneman (1979) have noted the need to empirically compare the various methods. The present study was conducted to answer the following questions: To what extent do the three methods (i.e. transformed item difficulty, Mantel-Haenszel, and b-parameter difference) methods agree or disagree in detecting a gender-related DIF? A second question was: What is the content or nature of those items identified as indicating DIF?
Samples

The samples used in the present study included 1600 males and 1790 females drawn from 11th grade students in Malaysia.

Instrument

The test of 45 dichotomous items assessing mathematical proficiency in four major areas: Basic arithmetic, verbal arithmetic, elementary algebra and geometry was developed by the author. In the unidimensional analysis, all 45 dichotomous items were regarded as measuring a single dimension by the use of factor analysis. Models-data fit investigated, 5 of 45 dichotomous items would be regarded as poorly fit by one parameter logistic model. The final version of the test comprised 40 items. For data analysis, SPSS, BILOG-MG, and Microsoft Excel were used.

DIF Methods

Three methods of detecting DIF (i.e. transformed item difficulty, M-H, and item-characteristic curve: b-parameter difference) were used to detect a gender-related DIF.

Results and Discussion

Data from the test was analyzed using the three previously described methods of detecting DIF (transformed difficulty, M-H, and item-characteristic curve: b-parameter difference) for the male and female groups. One pervasive question in bias research is whether differences are due to ability or to bias. It might be argued that an analysis should be based on male and female groups matched on ability (Lord, 1980). The basic premise is that a search for bias is one in which we seek to find items functioning differently for persons of the same ability in different groups. While some of the methods
attempt to control for ability differences, artifacts are still introduced because
the distributions of ability are markedly different in the two groups to be
compared (Subkoviak et al, 1987).

Hunter (1975) examines how this may cause unbiased items to be
identified as biased. One way to insure that the ability distributions are the
same is to match groups on ability. It should also be stressed that all of the
above bias methods are internal methods. That is, they all assume that the
average item is unbiased, and will not detect item bias if there is a constant
bias across items (Petersen, 1977). These methods will be most appropriate
during test development.

The M-H procedure flagged seventeen or 42.5 percent of the forty items
as indicating DIF (seven items were in favor of females and ten items were in
favor of males). The range of odds ratio signifies DIF in favor of males were
from 1.44 to 2.58, whereas the range of odds ratio signifies DIF in favor of
females were from .41 to .71.

Twenty-five or 62.5 percent of the forty items were easier for males (i.e.
the lowest value of b-parameter for one group indicates that the item is easy
for this group). The range of b-parameter difference signifies DIF in favor of
males were from .28 to 1.77, whereas the range of b-parameter difference
signifies DIF in favor of females were from -0.93 to -0.42. According to b-
parameter difference approach criteria, twenty-five or 62.5 percent of the forty
items revealed DIF (forteen items were in favor of males and eleven items
were in favor of females).

The TID procedure flagged sixteen or 40 percent of the forty items as
revealing DIF (one item was in favor of female students and the fifteen items
were in favor of male students). The range of D signifies DIF in favor of
males were from -1.73 to -1.0I, whereas the significant value of D for female was 1.97.

In order to inspect the consistency between any two of the three approaches in detecting DIF for the test, the percentage of pairwise agreements among the three approaches were computed (i.e. the degree of correspondence among each of pairwise methods with respect to the items revealing or not revealing the DIF for all items were computed).

M-H and b-parameter difference methods were agreeable in allocating nine items as revealing DIF, and thirteen items as not revealing DIF. As such, the percentage of agreement between M-H and b-parameter difference methods is 55% (i.e. 9 +13/40=55%) (See table 1).

**Table 1**

**Pairwise Agreement between M-H and b-parameter Difference Methods**

<table>
<thead>
<tr>
<th>Results From b-parameter difference</th>
<th>Results From M-H</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of nonflagged items</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Marginal total</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

Further, TID and M-H methods were agreeable in allocating seventeen items as revealing DIF, and nine items as not revealing DIF. As such, the percentage of agreement between M-H and TID methods is 65% (i.e. 17+9/40=65%) (See table 2).

**Table 2**

**Pairwise agreement between M-H and TID methods**

<table>
<thead>
<tr>
<th>Results From M-H</th>
<th>Results From TID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Marginal total</td>
<td>24</td>
<td>16</td>
</tr>
</tbody>
</table>
The TID and b-parameter methods were agreeable in allocating seven items as revealing DIF, and ten items as not revealing DIF. As such, the percentage of agreement between TID and b-parameter difference methods is 43% (i.e. 7+10/40=43%) (See table 3).

Table 3
Pair wise agreement between M-H and b-parameter difference methods.

<table>
<thead>
<tr>
<th>Results From b-parameter difference</th>
<th>Results From TID</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td>7</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>17</td>
</tr>
<tr>
<td>Marginal total</td>
<td>24</td>
</tr>
<tr>
<td>No. of Non flagged</td>
<td>6</td>
</tr>
<tr>
<td>No. of flagged</td>
<td>10</td>
</tr>
<tr>
<td>Marginal Total</td>
<td>16</td>
</tr>
</tbody>
</table>

In summary, the percentage of agreement among the three approaches in detecting DIF is relatively low. The range is from 43% to 65% for detecting DIF. The highest agreement was among M-H and TID methods, however, this may due to: both of these methods related to classical test theory. The lowest agreement was among TID and b-parameter difference. It does, of course, make sense for the transformed difficulty, M-H, and item b-parameter difference approaches to be in agreement since they all attempt to control for mean differences of the groups being compared and then concentrate on measuring relative difficulty. The b-parameter difference method should be least sensitive to distributional differences by virtue of the sample invariant quality of the parameters and the equating procedure. Both the M-H, which is in effect a comparison of observed item characteristic curves, and the Angoff transformed difficulty procedure, should be similar because they also adjust for ability as measured by total score.

The theoretical reasons for the lack agreement between certain pairs of methods in the identification of DIF of items are given by Hunter (1975). He
discusses several factors which may cause an item to be labeled as revealed DIF when, in fact, no DIF exists. These are (a) non-unidimensional tests, (b) differences in ability distribution of the two groups, (c) differences in item quality, (d) guessing, and (e) nonlinearity of regression. Finally, one should consider the fairness of an item in addition to its statistical index of bias. Also, this result helps to explain the low and moderate agreement reported in the measurement literature among DIF methods concerning items flagged as reveal DIF. The fact is that studies of convergence of methods for investigating DIF are influenced greatly by the unreliability of the statistics.

To better understand lack of agreement between certain pairs of methods, a closer look was taken at those items identified as revealing DIF by one method but not another. Rudner (1977) found that all the items identified as revealing DIF by both the transformed difficulty and b-parameter difference methods were items where the discrimination parameters were similar, but the location of the parameters differed.

The percentage of agreement across the three methods in detecting DIF was 15%. The items: 5, 8, 19, 20, 34, and 39 revealing DIF by the norms of the three methods. Only the items 20 and 39 were in favor of males students.

In previous studies, after a statistical procedure had been used to identify potentially biased items, attempts were made to identify possible content sources of this bias. In general, this procedure has neither provided consistent nor easily to generalize the results.

Mathematics items indicating DIF in favor of males were found to involve geometry and items involving real world references. Mathematics items indicating DIF in favor of females tended to involve algebra and miscellaneous regular mathematics items.
For mathematics items revealing DIF in favor of male student, the content characteristics involved (a) multiplication of decimal fractions, and (b) the use of more abstract mathematical concepts, such as, algebraic concepts, mathematical definitions, and the expression of a number in expanded notation.

Both mean raw scores and DIF index point to the conclusion that women had an advantage over men on the Algebra (item number 27), whereas men had an advantage on items involving proportional reasoning, Number and Computation, Data Analysis, and Geometry and Measurement. The tendency for men to perform better than women on Geometry and Measurement and women to perform better on Algebra is consistent with previous findings. In previous studies, however, women usually performed better on Number and Computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance. Such a strong female advantage in algebra, as reflected in DIF indexes has not been previously noted. The advantage could perhaps be explained by noting that the algebra items were very abstract and algorithmic, unlike the items in the other component, the algebra component included no real-world situation, no geometry, no word problems. As already noted, women tend to do better on such problems. Perhaps these gender related differences in performance are a result of both a reliance on routines learned in class as proposed by Bohlin (1994), Linn and Hyde (1989), Kimball (1989), and Gallagher (1992), and a lack of confidence on non routine tasks as suggested by Seegers and Boekaerts (1996).
References


Chapter Eight
Detecting DIF Using Logistic Regression and Transformed Item Difficulty

Abstract
The purpose of this study was to examine gender differences in performance on multiple-choice mathematical ability test, designed to match six grade curriculums. The LR (logistic regression) method and transformed item difficulty were used to detect a gender related DIF. A random sample of 800 tenth grade students was selected. DIF analysis indicated that: (1) Females showed a statistically significant and consistent advantage over males on numerical ability, whereas men showed a consistent advantage over females on spatial ability and deductive ability; (2) The percentage of agreement between the two approaches in detecting DIF is relatively low; and (3) Gender differences in mathematics may well be linked to content.

Introduction
Standardized tests and measurements are used primarily to distinguish between ability levels of examinees. As a part of the determination of validity for these tests, differential item analysis is employed to evaluate the degree to which measurements distinguish true abilities among examinees in an unbiased manner.

Psychometricians and test developers use DIF (differential item functioning) analysis to determine if there is a possible bias in a given test item. DIF is determined in a two-step process. The first step is the comparison of two groups’ outcome on an item and determining the presence of DIF. The
second step includes a decision of whether there is a large enough difference between the groups to eliminate or change the item of interest.

DIF is said to be present when examinees from different groups have differing probabilities of success on an item after controlling for overall ability (Clauser & Mazor, 1998). If an item is free of bias, responses to that item will be related only to the level of the underlying trait that the item is trying to measure. If item bias is present, responses to the item will be related to some other factors as well as the level of the underlying trait (Camilli & Shepard, 1999).

The tight relationship between the probability of correct responses and ability or trait levels is an explicit assumption of IRT (item response theory) (Hambelton, Swaminathan, & Rogers, 1991) and an implicit assumption of classical test theory (McDonalds, 1999).

The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF.

Test items with content bias may: (1) contain content that is differentially familiar to matched groups of examinees; (2) contain sources of difficulty that are irrelevant to the construct adversely affecting test performance; (3) contain material that may be offensive, demeaning, or emotionally charged which can lower examinees’ motivation and attention for the remainder of the test, thereby, decreasing performance on other questions apart from the offending items; and (4) ask for information that students have not had equal opportunity to learn. Test items with gender bias may contain: (1) tasks which perpetuate undesirable role stereotypes, race stereotypes or gender stereotypes; (2) materials or references that may be offensive to
members of one gender; and (3) references to objects and ideas that are likely to be more familiar to men or to women (Pedrajita, 2009).

Several techniques have been promulgated for the statistical assessment of DIF. Several excellent reviews are available (Clauser & Mazor, 1998; Camilli & Shepard, 1999; Millsap, 1993). Most techniques for DIF assessment has been developed in educational settings in which items are generally dichotomously scored as correct or incorrect. A number of approaches have used item difficulty as the focus of analysis. An item is considered biased in this approach if, compared to other items on the test, it is relatively more difficult for one group than for another.

One of the more widely implemented techniques of this type is TID (transformed item difficulty). LR (Logistic Regression) is based on transforming data by taking their natural logarithms so as to reduce nonlinearity. In other words, LR uses the logistic curve that best approximates the distribution of the data. LR estimates parameters using maximum likelihood estimation (Pedrajita, 2009). LR has been known for some time to be useful for the assessment of effect modification in observational studies and enables analyses of continuous predictor variables without requiring stratification. Not surprisingly, simulation studies from educational testing experts have found that LR-based DIF detection techniques enable the detection of both uniform and non-uniform DIF.

**Gender Differences in Mathematics**

In the past few decades, research has repeatedly reported gender differences in mathematics performance on a number of standardised mathematics tests such as the SAT-M (Scholastic Assessment Test-
Mathematics) (Gallagher, 1990, 1992; Gallagher & DeLisi, 1994; Willingham & Cole, 1997; Hyde, Royer et al, 1999). The test scores on these standardized tests have been regarded as an important measure of abilities to do mathematics problems (Casey, Nuttall, Pezaris, & Benbow, 1995; Halpern, 2000; Stumpf & Stanley, 1998). But results from these studies are not consistent: Some found that males generally outperformed females on mathematical tasks (Maccoby & Jacklin, 1974; Fennema & Carpenter, 1981; Halpern, 2000); some showed different sizes of gender differences with respect to types of mathematical tasks (D. Voyer, S. Voyer, & Bryden, 1995). Hyde, Fennema and Lamon (1990) suggested that there was very small or null gender difference in mathematical ability on these tests. T. B. Caplan and P. J. Caplan (2005) even argued that the link between gender and the mathematical ability was very weak.

Battista (1990) conducted a study among 145 high school geometry students from middle-class communities. This research examined the role that spatial visualization and verbal-logical thinking played in gender differences in geometric problem-solving in high school. The findings suggested that males and females differed in the level of discrepancy between spatial and verbal abilities.

Gallagher et al (2000) suggested that males tended to be more flexible than females in applying solution strategies. Kessel and Linn (1996) and Gallagher (1998) reported that females were more likely than males to adhere to classroom-learned procedures to solve problems, so they might be less likely to use shortcuts and estimation techniques for solving unfamiliar and complex problems quickly.
This study provided an opportunity to examine issues in mathematics learning in general and issues in gender-related differential item functioning of mathematical ability in specific.

**Purpose**

This study aimed to detect DIF of mathematics ability test. This study can significantly contribute to educational research. Test experts and developers may: (1) gain insights on the applicability of DIF detection method(s); (2) realize the validity of DIF methods in detecting a gender biased test items; (3) use DIF methods in developing valid and equitable tests; and (4) employ DIF methods in purifying their assessment instruments.

This study sought answers to the following questions: To what extent do the two methods (i.e., transformed item difficulty and LR) agree or disagree in detecting a gender-related DIF? A second question was: What is the nature of cognitive ability of those items identified as revealing DIF? A third question was: Are gender differences linked to content areas within mathematics?

**Participants**

A total of 800 (380 males and 420 females Grade 10) Jordanian students were targeted as participants in this study at the end of the first semester in the school year of 2009-2010.

**Instrument**

A mathematical ability scale was developed as a part of this study. The scale compressed of 40 multiple-choice items to measure three components of mathematical ability (i.e., numerical ability, deductive ability and spatial ability). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported $KR$ (Kuder-Richardson)-20.
indices to be .91. Spearman-Brown correction on split-half reliability for odd even comparison also show similar results $r = .89$. Validity of the instrument was shown using inter-correlation of the scale (.19 to .855).

**DIF Methods**

Two methods (i.e., transformed item difficulty (TID) and Logistic Regression (LR)) were used for DIF detection.

**Results and Discussion**

LR method was used to identify DIF on the mathematical ability scale for each of 30 items. Sixteen items or 43% of the items revealed DIF (eight items were revealed uniform DIF and eight items were revealed non-uniform DIF). Ten items were in favor of males and six items were in favor of females.

Furthermore, the TID method flagged ten items at the significance level of .05 (one item was in favor of female students, whereas nine items were in favor of male students).

The two methods were agreeable in allocating 14 items as revealing DIF, and ten items as not revealing DIF. As such, the percentage of agreement between TID and LR methods is 45% (i.e., $16 + 2/40 = 45\%$) (See table 1).

<table>
<thead>
<tr>
<th>Results From TID</th>
<th>Results From LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td>16</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>14</td>
</tr>
<tr>
<td>Marginal total</td>
<td>30</td>
</tr>
</tbody>
</table>

In summary, the percentages of agreement among the two approaches in detecting DIF are relatively low. Not surprisingly, simulation studies from educational testing experts have found that LR-based DIF detection techniques
enable the detection of both uniform and non-uniform DIF, whereas TID DIF
detection techniques unable the detection of both non-uniform DIF.

The theoretical reasons for the lack agreement between both methods in
the identification of DIF of items are given by Hunter (1975) who discussed
several factors which may cause an item to be labeled as revealed DIF when,
in fact, no DIF exists. These are: (1) non-unidimensional tests; (2) differences
in ability distribution of the two groups; (3) differences in item quality; (4)
guessing; and (5) nonlinearity of regression. Finally, one should consider the
fairness of an item in addition to its statistical index of bias. Also, this result
helps to explain the low and moderate agreement reported in the measurement
literature among DIF methods concerning items flagged as reveal DIF. The
fact is that studies of convergence of methods for investigating DIF are
influenced greatly by the unreliability of the statistics (Abedalaziz, 2010).

The DIF analysis pointed to the conclusion that females had an
advantage over males on the numerical ability, whereas males had an
advantage over females on items involving spatial ability and deductive
ability. The tendency for males to perform better than females on spatial
ability and inductive ability and women to perform better on numerical ability
is consistent with previous findings (Willson, Fernandez, & Hadaway, 1993;
Gallagher et al, 2000).

In previous studies, however, females usually performed better on
number and computation. The fact that this test was tied to a specific
curriculum did not appear to help females’ performance. The researchers
consistently found that male students are superior in geometry and
visualization (Geary, 1996). On the other hand, females show superiority in
computation based on the data available.
Spatial abilities were reported to have relationship with mathematics test scores (Casey, Nuttall, Pezaris, & Benbow, 1995; Geary, Saults, Liu, & Hoard, 2000; Nuttall, Casey, & Pezaris, 2005). This relationship indicates that gender differences in spatial abilities may contribute to gender differences in mathematical problem-solving.

The study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum. The presence of a gender related DIF in mathematical ability test can be attributed to: (1) the unfamiliar with the content of the items which caused the examinees to be attracted to the incorrect options; (2) the ambiguities in the item stem, keyed response, or distracter; (3) the disparities in the matched examinees’ exposure to concepts or skills reflected on the items; and (4) the inability of the matched examinees to understand the concepts reflected on the items (Pedrajita, 2009).

References

161


Chapter Nine

Detecting DIF using Item Characteristic Curve Approaches

Abstract

In the event an item is functioning differently for one of the groups, a decision must be made about whether to retain the item or delete it from the test. If the item proves to be biased against a subgroup, its magnitude is strong enough to bias test results, and a rationale exists for why it may exhibit DIF, then the item should be deleted (Camilli & Shepard, 1994). However, without a substantive review of the item to understand the reason it resulted in DIF, test developers do not actually know if the source of DIF is due to a construct-relevant or irrelevant dimension being measured by the test. Therefore, it is important to accurately interpret the nature of DIF so that differences between the groups’ cognitive skills or opportunities to learn can be appropriately addressed. The present study sought answers to the following questions: (1) to what extent does the two methods (i.e. area index for three-parameter logistic model, and area index for two parameter logistic model) agree or disagree in the identification DIF? (2) Are there gender differences in mathematical proficiency? (3) Are gender differences linked to content areas within mathematics? In order to answer these questions, achievement test of tenth grade covering the following subjects: Relations and functions, polynomial, Trigonometric functions, and triangles were developed, and administered to samples of 1318 tenth grade students (664 males and 654 females). Results indicated that: (1) the percentage of agreement between the two approaches in detecting DIF is relatively high. The range is from 81% to 89% for detecting uniform and non uniform DIF respectively. (2) There is a gender-related DIF
linked to content areas within mathematics. (3) Availability of gradual differential functioning in a number of dichotomous items with respect to the nature of mathematical context being measured by the two approaches which was non uniform in general.

**Introduction**

Item bias and differential item functioning (DIF) have critical political, social, and ethical implications for tests developers, policy makers, and examinees. The study of item bias and DIF is critical; as such research helps provide an empirical foundation for the identification and subsequent elimination of exam items that appear to be relatively more difficult for one group of test-makers than another. Further research on these issues will allow us to comprehend more fully the possible substantive interpretation that can be made by focusing upon test items considered to be biased. In addition, subsequent research can help us understand in greater depth the factors that contribute to DIF.

Gender related differential item functioning is a constant concern on large-scale standardized achievement tests in mathematics because differences between females and males are often found (e.g., Bielinski & Davison, 2001; Boughton, Gierl & Khalaq, 2000; DeMars, 1998; Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997; Abedalaziz, 2010). Presumably, because of the complexity of gender-related issues, results reported from a variety of studies are inconsistent and often even contradictory (Cleary, 1992; Hyde, 1991; Willingham & Cole, 1997; Abedalaziz, 2010) suggest that such contradictory results may be accounted for by disentangling effects of different cohorts, construct, and selectivity of the sample.
**Research Questions**

The present study sought answer to the following question: to what extent do the two methods (i.e. area index for three-parameter logistic model, and area index for two parameter logistic model) agree or disagree in the identification of a gender-related DIF?

**Samples**

In total, 1318 tenth grade Jordanian students (664 males, and 654 females) participated in the study.

**Instrument**

A test compressed 40 dichotomous items assessing mathematical proficiency in four major areas: relations and functions, polynomial, trigonometric functions, and triangles were developed by the author. In the unidimensional analysis, all 40 dichotomous items were regarded as measuring a single dimension. 3 of 40 items were poorly fit the models, and discarded.

**Data Analysis**

The three-parameter logistic model and the two-parameter logistic models were fitted to the data, item parameters were estimated with the BILOG (version 3; Mislevy & Bock, 1990) program, and the Expected A Posteriori (EAP) scoring method was used within the BILOG program to estimate θ. Area differential index procedure for 3- parameter Logistic model, and area differential index procedure for 2- parameter Logistic model were used to explore a gender-related DIF. Microsoft excel was used to calculate the area between the two curves using Raju equations ( Raju, 1988). The critical areas (cut-off values) for the two approaches were determined (i.e. the
critical area for 3-parameter logistic model was 0.0222, whereas the critical area for 2-parameter logistic model was 0.0333). In the present study, the item reveals DIF when the area between the two curves is greater than the cut-off area.

Since the c parameter for the two groups are not the same, the significance test for the area statistic cannot be carried out. The problem is to find "cut-off" values for the area statistic that can be used to decide whether DIF is present. An empirical approach to determining a cut-off is to divide the group with the larger sample size into two randomly equivalent groups, to estimate the ICC_S in each group separately, and to determine the area between the estimated ICC_S (Hambleton & Rogers, 1989). Since the groups are randomly equivalent, the area should be zero. Nonzero values of the area statistic are regarded as resulting from sampling fluctuations, and the largest area value obtained may be regarded as the largest value that may be expected in terms of sampling fluctuation. Any area value greater than this is assumed to be "significant" and, consequently, indicative of DIF when the focal and reference groups are compared.

**Results and Discussion**

Item difficulty (b), item discrimination (a), item guessing (c), and the area between the two curves for each of the thirty-seven items were extracted. The area index for the 3-parameter logistic model procedure flagged twenty-eight or 76 percent of the thirty-seven items as revealing DIF. The range of significant area is from .03 to .3973.

The visual inspection of the ICCs graphs will reveal whether the DIF is uniform or nonuniform. The item characteristics curves for 3-parameter
logistic model were obtained. Tenth or 27 percent of the thirty-seven items are revealing uniform DIF (five items were in favor of males, and five items were in favor of females). Sixteen or 43 percent of the thirty-seven items are revealing nonuniform DIF. For example, the item 12 reveals DIF in favor of males from -3 to 0 logit, whereas in favor of females from .0 to 3.0 logit (Figure 1). A look to the curves, two items were not revealing DIF across all levels of ability.

The item difficulty (b), item discrimination (a), and the area between the two curves for each of the thirty-seven items were extracted. The area index for 2-parameter logistic model procedure flagged twenty-nine or 78 percent of the thirty-seven items as indicating DIF. The range of significant area was from 0.05 to 1.56. Eight or 22 percent of the thirty-seven items revealed uniform DIF (three items were in favor of males and five items were in favor of females). Seventeen or 46 percent of the thirty-seven items revealed nonuniform DIF. For example, the item 16 reveals DIF in favor of
males from -3 to -0.5 logit, whereas in favor of females from -0.5 to 3 logit (Figure 2). A look to the curves, the items 14, 31, and 35 are not revealing DIF across all levels of ability.

![Graph showing item (16) revealing uniform DIF](image)

**Figure 2: Item (16) revealing uniform DIF**

In order to inspect the consistency between the two approaches in detecting DIF of a mathematical ability scale, the percentage of agreements among the two approaches was computed (i.e. the percentage of items revealing or not revealing DIF).

The percentage of agreement between the area index for 3-parameter logistic model and the area index for 2-parameter logistic model approaches is 81% (i.e. 5+25/37 =81%) (See table 1). The percentage of agreement between the two approaches in detecting uniform DIF is 81% (i.e. the items: 13, 17, 22, 27, 34, and 39 are revealing uniform DIF), whereas the percentage of agreement between the two approaches in detecting non uniform DIF is 78%
(i.e. the items: 4, 7, 9, 12, 15, 16, 21, 23, 25, 26, 28, 32, and 40 are revealing non uniform DIF) (see table 2).

Table 1
The agreement between the two area indexes

<table>
<thead>
<tr>
<th>Results From the area index for 3-Parameter approach</th>
<th>No. of Non flagged</th>
<th>No. of flagged</th>
<th>Marginal Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of non flagged items</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>4</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>Marginal total</td>
<td>9</td>
<td>28</td>
<td>37</td>
</tr>
</tbody>
</table>

Table 2
The agreement between the area index for 3-parameter logistic model and the area index for the 2-parameter logistic model approaches in detecting Uniform and Non uniform DIF

<table>
<thead>
<tr>
<th>Results From the area index for 3-Parameter approach</th>
<th>Uniform</th>
<th>Non uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of nonflagged Items</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Marginal Totals</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>No. of nonflagged Items</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>No. of flagged Items</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Marginal Totals</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Marginal Totals</td>
<td>37</td>
<td>37</td>
</tr>
</tbody>
</table>

In summary, the results indicated that: (1) Availability of gradual differential functioning in a number of items with respect to the nature of mathematical context being measured by the two approaches which was non uniform in general. However, this may due to some factors such as: mutuality, test fairness, experience, and guising. (2) The percentage of agreement between the two approaches in detecting DIF is relatively high. The range is from 81% to 89% for detecting uniform and non uniform DIF, respectively. However, this may due to both methods related to Item Response Theory. (3) Some of the items revealing differential item functioning by one of the
approaches did not reveal the same by the other approach; however, this may
due to the experimental approach of determining a cut-off area. This approach
is not ideal; however, it does provide an approximate answer to the cut-off-
score determination problem (Hambleton, Swamanithan, & Rogers, 1991).

IRT-based methods somewhat unreliable in identifying differentially
functioning items. Finally, Hunter (1975) notes several flaws in the item
characteristic curve methods. Differences in the ability distributions will be
reflected in instability at the ends of the curves for different groups and
different displacements of the item-characteristic curves because of
unreliability. He finally states that all methods fail if the test is not
unidimensional.

Prieler and Raven (2009) summarized five problems involved in
measuring the problem of assessing differential change at different levels of
ability: (a) Problems arising from floor and ceiling effects. (b) Problems
arising from the frequently encountered need to use a different and more
difficult test to assess performance after an intervention. (c) Problems arising
from the available tests not yielding equal-interval scales. (d) Problems arising
from a preoccupation with single-variable assessments. (e) Problems arising
from the low reliability, construct validity, and predictive validity – and thus
meaningfulness – of differences between individual scores before and after
some treatment.

Several major points emerge from these finding. First, IRT-based
methods somewhat unreliable in identifying differentially functioning items.
This finding reinforces our preference for considering items only "potentially
biased" on the basis of the value of the DIF statistic.
Also, this result helps to explain the moderate agreement reported in the measurement literature among DIF method concerning item flagged as potentially biased. The fact is that studies of convergence of methods for investigating DIF are influenced greatly by the unreliability of the statistics. Second, in some items, the DIF was outstanding in the highest level of ability and standing in the lowest level of ability, and in other items the DIF was outstanding in the lowest level of ability and standing in the highest level of ability.

In conclusion, note that the results of this study warrant further investigation before too many generalizations can be made. Future follow up studies in which item formats and content are experimentally manipulated would enhance our understanding of these variables. Also, note that (over one half) of the categories that were considered failed to show gender differences between groups of male and female students matched on overall test scores. However, an examination of those that did show differences yield a better understanding of more global differences in test performance. The fact that this test was tied to a specific curriculum did not appear to help females' performance (for example, Maccoby & Jacklin, 1974; Fennema & Carpenter, 1981; Halpern, 2000).

For those five items identified by area index for 3-parameter logistic model as revealing DIF in favor of females, were verbal application items. The five items identified as revealing DIF in favor of males, were in scientific and mathematical tasks. Also, the five items identified by area index for 2-parameter logistic model as revealing DIF in favor of females, were verbal application items and items resembling textbook. The three items identified as
revealing DIF in favor of males, were in scientific and mathematical task, and nonconventional.

References


Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 12(6), 18-34.
Chapter Ten

Detecting DIF using Logistic Regression

Abstract

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1990). When personal attributes, such as gender systematically affect examinee performance on an item, the result can be differential item functioning (DIF). The logistic regression procedure for Differential Item Functioning (DIF) detection is a model-based approach designed to identify both uniform and nonuniform DIF. The purpose of this study was to examine gender differences in performance on multiple-choice mathematical ability test, designed to match six grade curriculum. The logistic Regression method was used to detect a gender related DIF. A random samples of 800 tenth grade students were selected. In DIF index, females showed a statistically significant and consistent advantage over males on numerical ability, whereas men showed a consistent advantage over females on spatial ability and deductive ability.

Introduction

Bias is a serious problem in psychometric tests. Differential item functioning (DIF) is said to be present when examinees from different groups have differing probabilities of success on an item, after controlling for overall ability (Clauser and Mazor, 1998). If an item is free of bias, responses to that item will be related only to the level of the underlying trait that the item is
trying to measure. If item bias is present, responses to the item will be related to some other factor as well as the level of the underlying trait (Camilli and Shepard, 1999) The tight relationship between the probability of correct responses and ability or trait levels is an explicit assumption of item response theory (IRT) (Hambelton et. al., 1991) and an implicit assumption of classical test theory (McDonalds, 1999) The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF.

Several techniques have been promulgated for the statistical assessment of DIF. Several excellent reviews are available (Clauser and Mazor, 1998; Camilli and Shepard, 1999; Millsap, 1993). Most techniques for DIF assessment were developed in educational settings in which items are generally dichotomously scored as correct or incorrect. It was recognized by the early 1990s that LR-based techniques were more powerful than MH-based techniques (Jodin and Gierl, 2001; Rojers & Swaminthan, 1993; Swaminathan & Rojers, 1990). This power may come at the expense of increased type I error rates in LR-based techniques (Jodin and Gierl, 2001).

Two distinct forms of DIF have been recognized. These have been called uniform and non-uniform DIF. Uniform DIF is said to apply when differences between groups in item responses are found at all trait levels, while in non-uniform DIF an interaction is found between trait level, group assignment, and item responses (Camilli and Shepard, 1999; Jodin and Gierl, 2001) LR has been known for some time to be useful for the assessment of effect modification in observational studies, and enables analyses of continuous predictor variables without requiring stratification. Not surprisingly, simulation studies from educational testing experts have found
that LR-based DIF detection techniques enables the detection of both uniform and non-uniform DIF.

**Gender differences in mathematics**

In the past few decades, research has repeatedly reported gender differences in mathematics performance on a number of standardised mathematics tests such as the Scholastic Assessment Test-Mathematics (SAT-M) (Gallagher, 1990, 1992; Gallagher and DeLisi, 1994; Hyde, Fennema, and Lamon, 1990; Royer, Tronsky, Chan, Jackson and Marchant, 1999; Willingham and Cole, 1997). The test scores on these standardised tests have been regarded as an important measure of abilities to do mathematics problems (Casey, Nuttall, Pezaris, and Benbow, 1995; Halpern, 2000; Stumpf and Stanley, 1998). But results from these studies are not consistent: some found that males generally outperformed females on mathematical tasks (for example, Maccoby and Jacklin, 1974; Fennema and Carpenter, 1981; Halpern, 2000); some showed different sizes of gender differences with respect to types of mathematical tasks (for example, Voyer, and Bryden, 1995). Hyde, et al. (1990) suggested that there was very small or null gender difference in mathematics performance on these tests. Caplan and Caplan (2005) even argued that the link between gender and the mathematics performance was very weak. Can test scores measure the real differences in cognitive abilities and abilities to solve mathematical problems between females and males?

Gender differences were evident in successful patterns and in strategy use on conventional and unconventional problems…female students were more likely than male students to correctly solve “conventional” ³ problems (by) using algorithmic strategies; male students were more likely than female
students to correctly solve “unconventional” problems (by) using logical estimation and insight. (Gallagher et al., 2000, p.167).

Battista (1990) conducted a study among 145 high school geometry students from middle-class communities. This research examined the role that spatial visualisation and verbal-logical thinking played in gender differences in geometric problem solving in high school. The findings suggested that males and females differed in the level of discrepancy between spatial and verbal abilities.

Gallagher et al. (2000) suggested that males tended to be more flexible than females in applying solution strategies. Kessel and Linn (1996) and Gallagher (1998) reported that females were more likely than males to adhere to classroom-learned procedures to solve problems, so they might be less likely to use shortcuts and estimation techniques for solving unfamiliar and complex problems quickly.

Current education reform in general and mathematics education reform in particular emphasize the importance of thinking, understanding, reasoning, and mathematical ability in students' learning (e.g., NCTM, 1989, 1991, 2000; National Research Council, 1989). Such reform effort in mathematics curriculum and instruction requires examination of male and female students' thinking, reasoning, problem solving and mathematical ability rather than merely computation and symbol manipulation. This study provided an opportunity to examine issues in mathematics learning in general and issues in gender – related differential item functioning of mathematical ability in specific.

**Purpose**
This study sought answers to the following questions: Are there gender differences in mathematical ability? Are gender differences linked to content areas within mathematics?

**Method**

**Participants**

A total of 800 (380 males and 420 females) tenth grade students in Jordan were targeted as participants in this study, during the ending period of the First semester, school year 2009-2010.

**Instrument**

A mathematical ability scale was developed as a part of this study. The scale compressed of 30 multiple-choice items to measure three components of mathematical ability (i.e. numerical ability, deductive ability, and spatial ability). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be 0.91. Spearman-Brown Correction on split-half reliability for odd even comparison also show similar results r=0.89. Validity of the instrument was shown using inter-correlation of the scale (0.19 to .855). Factor Analysis reveals that the test measure one trait (unidimensionality).

**Detecting DIF**

Methods for detecting DIF have proliferated in recent years and have been reviewed. The various methods include techniques that tested differences in relative item difficulty among different groups, differences in item discrimination among different groups, differences in the item-characteristic curves (ICCs) for different groups, differences in the distribution of incorrect
responses for various groups, and differences in multivariate factor structures among groups. In the present study, Logistic Regression approach used to detect DIF.

Swaminathan and Rogers (1990) applied the Logistic Regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and nonuniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and nonuniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of nonuniform DIF compared to the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). These studies also identified two major weaknesses in the LR DIF procedure: 1) the Type I error or false positive rate was higher than expected, and 2) the lack of an effect size measure.

Logistic regression has a formal mathematical equivalence to the log linear model approach of Mellenbergh (1982): Coefficients for group, total score, and interaction terms are estimated and tested for significance with a model comparison strategy. However, logistic regression is highly similar to standard ordinary least squares regression. It can be conceptualized as an equation that uses group, ability, and group-by-ability terms to predict whether an item response is right (1) or wrong (0). This property is desirable for didactic purposes.

In the present study, the item reveals uniform DIF when the significant odd ratio is for the group, whereas the item reveals non uniform DIF when the
significant odd ratio is for the interaction between the group and total score. The item reveals DIF in favor of males when the significant odd ratio is greater than one, whereas the item reveals DIF in favor of females when the significant odd ratio is less than one ($\alpha=0.05$).

Results

Tables 1 shows the summary results of the Logistic Regression method to identify Differential Item Functioning on the mathematical ability scale for each of the thirty items. Seventeen items or 57 percent of the items revealed DIF (i.e. the items: 1, 5, 8, 21, 22, 23, 24, and 26 were revealed uniform DIF, whereas the item: 9, 10, 11, 13, 16, 27, 28, and 29 were revealed non uniform DIF). The items: 1, 8, 10, 13, 16, 21, 24, 26, 28, and 29 were in favor of males, whereas the items: 5, 9, 11, 22, 23, and 27 were in favor of females).
<table>
<thead>
<tr>
<th>Item</th>
<th>Variable</th>
<th>Statistical significance</th>
<th>Odds-Ratio</th>
<th>Type of DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group</td>
<td>0.0146</td>
<td>2.6975</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1106</td>
<td>0.9361</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Group</td>
<td>0.1693</td>
<td>1.7383</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.3355</td>
<td>1.0433</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Group</td>
<td>0.7394</td>
<td>0.8146</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1895</td>
<td>1.0547</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Group</td>
<td>0.6374</td>
<td>1.2300</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0349</td>
<td>1.0110</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Group</td>
<td>0.0001</td>
<td>0.1410</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0601</td>
<td>1.2605</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Group</td>
<td>0.1399</td>
<td>0.3481</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0883</td>
<td>1.0833</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Group</td>
<td>0.7014</td>
<td>0.8516</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.5831</td>
<td>1.0257</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Group</td>
<td>0.0136</td>
<td>2.6964</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1103</td>
<td>0.9361</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Group</td>
<td>0.0046</td>
<td>0.2549</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0238</td>
<td>1.1367</td>
<td>Nonuniform</td>
</tr>
<tr>
<td>10</td>
<td>Group</td>
<td>0.6990</td>
<td>2.6234</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0072</td>
<td>0.8781</td>
<td>Nonuniform</td>
</tr>
<tr>
<td>11</td>
<td>Group</td>
<td>0.0800</td>
<td>1.1749</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0001</td>
<td>1.2538</td>
<td>Nonuniform</td>
</tr>
<tr>
<td>12</td>
<td>Group</td>
<td>0.1331</td>
<td>1.9886</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.4770</td>
<td>0.9655</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Group</td>
<td>0.0972</td>
<td>2.6889</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0214</td>
<td>0.9166</td>
<td>Nonuniform</td>
</tr>
<tr>
<td>14</td>
<td>Group</td>
<td>0.7044</td>
<td>0.8767</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.7405</td>
<td>0.9987</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Group</td>
<td>0.2468</td>
<td>0.5718</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.6688</td>
<td>0.9830</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Group</td>
<td>0.0387</td>
<td>2.1889</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0234</td>
<td>0.9874</td>
<td>Nonuniform</td>
</tr>
<tr>
<td>17</td>
<td>Group</td>
<td>0.0059</td>
<td>3.1315</td>
<td>Nonuniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0033</td>
<td>0.8967</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Group</td>
<td>0.4950</td>
<td>0.7738</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.6900</td>
<td>0.9870</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Group</td>
<td>0.1244</td>
<td>0.4939</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.7097</td>
<td>0.9139</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Group</td>
<td>0.7527</td>
<td>0.8720</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.6213</td>
<td>0.9828</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Group</td>
<td>0.0068</td>
<td>1.7702</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.5432</td>
<td>0.9951</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Group</td>
<td>0.0177</td>
<td>0.1328</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1285</td>
<td>0.0529</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Group</td>
<td>0.0303</td>
<td>0.3351</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1151</td>
<td>1.0351</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Group</td>
<td>0.0490</td>
<td>1.2586</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1227</td>
<td>0.5649</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Group</td>
<td>0.4476</td>
<td>1.9674</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.2477</td>
<td>0.9048</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Group</td>
<td>0.0490</td>
<td>1.2579</td>
<td>Uniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.1226</td>
<td>0.5648</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Group</td>
<td>0.1283</td>
<td>0.3759</td>
<td>Nonuniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0266</td>
<td>1.0739</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Group</td>
<td>0.0673</td>
<td>3.5217</td>
<td>Nonuniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0242</td>
<td>0.8809</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Group</td>
<td>0.0531</td>
<td>0.7740</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.6901</td>
<td>0.9880</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Group</td>
<td>0.1348</td>
<td>2.7924</td>
<td>Nonuniform</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.0344</td>
<td>0.9363</td>
<td></td>
</tr>
</tbody>
</table>
Discussion and Conclusion

The DIF index point to the conclusion that females had an advantage over males on the numerical ability, whereas males had an advantage on items involving spatial ability and deductive ability, The tendency for males to perform better than females on spatial ability and inductive ability, and women to perform better on numerical ability is consistent with previous findings (e.g. Willson, Fernandez and Hadaway, 1993; Gallagher, DeLisi, Holst, McGillicuddy-DeLisi, Morely and Cahalan, 2000).

In previous studies, however, females usually performed better on Number and Computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance The Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, female show superiority in computation based on the data available. Gender differences in achievement in mathematics in favor of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

There are many studies that focus on differences between men and women in tests (Gallaghe, De Lisi, Holset, Mc Gillicuddy-De Lisi, Morly & Cahalan, 2000; Kimball, 1994; Willingham & Cole, 1997). From the findings of earlier studies, one conclusion can be drawn is that men have a better spatial ability than women (Geary, 1996). Men use this spatial more often than women when solving problems, which can give advantages while solving
certain kinds of problems in geometry (Geary, 1996). Many studies indicate that women are better than men in verbal skills, which can give them advantages on items where communication is important. Women also score relatively higher on tests in mathematics that better match coursework. Men tend to outperform women in geometry and in arithmetic and algebraic reasoning questions. Women tend to be better at intermediate algebra and arithmetic and algebraic operations (Willingham & Cole, 1997). Gallagher, De Lisu, Holset, Mc Gillicuddy-De Lisi, Morly & Cahalan, (2000) found men outperformed women in all kind of problems, but that the differences were greater for problems requiring spatial skills or multiple solution paths than for problems requiring verbal skills or containing classroom-based content.

Spatial abilities were reported to have relationship with mathematics test scores (Casey, Nuttall, Pezaris and Benbow, 1995; Geary, Saults, Liu, and Hoard, 2000; Robinson, Abbott, Berninger and Busse, 1996; in Nuttall et al, 2005). This relationship indicates that gender differences in spatial abilities may contribute to gender differences in mathematical problem solving.

The study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum.

References


McDonald RP. (1999). Test Theory: A Unied Treatment. Lawrence Erlbaum: Mahwah, NJ.


yes
I want morebooks!

Buy your books fast and straightforward online - at one of the world’s fastest growing online book stores! Environmentally sound due to Print-on-Demand technologies.

Buy your books online at
www.get-morebooks.com

Kaufen Sie Ihre Bücher schnell und unkompliziert online – auf einer der am schnellsten wachsenden Buchhandelsplattformen weltweit! Dank Print-On-Demand umwelt- und ressourcenschonend produziert.

Bücher schneller online kaufen
www.morebooks.de

OmniScriptum Marketing DEU GmbH
Heinrich-Böcking-Str. 6-8
D - 66121 Saarbrücken
Telefax: +49 681 93 81 567-9

info@omnisciptum.com
www.omnisciptum.com