Macrothink Institute™

International Journal of Education
ISSN 1948-5476
2012, Vol. 4, No. 2

Revisiting Differential Item Functioning: Implications for Fairness Investigation

Jinyan Huang** and Turgay Han*

***Associate Professor and Ph.D. Faculty Member*

College of Education, Niagara University

325F Academic Complex, Niagara University, NY 14109, United States

Tel: 1-716-286-8259 E-mail: jhuang@niagara.edu

**English Lecturer*

Faculty of Letters, Kafkas University, Turkey

E-mail: turgayhan@yahoo.com.tr

Received: February 3, 2012 Accepted: March 19, 2012 Published: April 17, 2012

doi:10.5296/ije.v4i2.1654 URL: http://dx.doi.org/10.5296/ije.v4i2.1654

Abstract

Fairness has been the priority in educational assessments during the past few decades. Differential item functioning (DIF) becomes an important statistical procedure in the investigation of assessment fairness. For any given large-scale assessment, DIF evaluation is suggested as a standard procedure by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. This procedure often affords opportunities to check for group differences in test performance and investigate whether or not these differences indicate bias. However, current DIF research has received several criticisms. Revisiting DIF, this paper critically reviews current DIF research and proposes new directions for DIF research in the investigation of assessment fairness.

Keywords: *differential item functioning; item bias; item impact; assessment fairness*

1. Introduction

Fairness has been the priority in educational assessments during the past few decades (Cole & Zieky, 2001). According to American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999), educational organizations, institutions, and individual professionals should make assessments as fair as possible for test takers of different races, genders, and ethnic backgrounds. Related to assessment fairness is the term Differential Item Functioning (DIF). Lord (1980) provided the following definition for DIF: If each test item has exactly the item characteristic curve in every group, then people of the same ability would have exactly the same chance of getting the item tight, regardless of group membership … If, on the other hand, an item has a different item characteristic curve for one group compared to another, it is clear the item is functioning differently across groups.

DIF research has received increased attention in educational and psychological contexts (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Dickinson, Wanichtanom, & Coates, 2003; Gierl, Rogers, & Klinger, 1999; Huang & Sheeran, 2011). For any given large-scale assessment, DIF evaluation is suggested as a standard procedure by AERA, APA, and NCME (1999). This procedure often affords opportunities to check for group differences in test performance and investigate whether or not these differences indicate bias.

However, many concerns have been raised about current DIF research. For example, there are several statistical methods for detecting DIF; but these methods do not yield consistent and stable results (Gierl et al., 1999). Further, many DIF research studies did not include the judgmental analysis (Camilli & Shepard, 1994). These concerns may limit the interpretation of the impact of DIF on test development. Therefore, it is important to revisit DIF and provide implications for research and practice in the area of investigating fairness in educational assessments.

Understanding the concept of DIF is the first step; the next is to discuss the procedures of detecting DIF; and finally it is to explore why DIF occurs and provide implications for fairness investigation in educational research. This paper first introduces some basic terms and their definitions. It then compares and contrasts various statistical procedures of detecting DIF. After that it summarizes current research results on the interpretation and explanation of DIF. In the following sections it criticizes current DIF research and proposes new directions for DIF research in the investigation of assessment fairness.

2. DIF, Item Bias, and Item Impact

DIF occurs when an item is substantially harder for one group than for another group after the overall differences in knowledge of the subject tested are taken into account. Therefore, DIF refers to the ways in which items function differently for individuals or groups of test takers with similar abilities (Kunnan, 1990). The DIF analysis is based on the principle of comparing the performance of focal groups (e.g., female, or Black examinees) on an item with that of a reference group (e.g., male or White examinees), in a way that controls for overall knowledge of the subject tested.

DIF does not mean simply that an item is harder for one group than for another; if the examinees in one group tend to know more than the other group about the subject, they will tend to perform better on all exam items. Therefore, once DIF is identified, it may be attributed to either item bias or item impact (Gierl et al., 1999; Huang & Sheeran, 2011).

Item bias, one potential threat to validity of the test, leads to systematic errors that misinterpret the inferences made for members of a particular group from a test. In other words, when an item unfairly favors one group other than another one, bias occurs. Items are biased because the items themselves contain certain sources of difficulty other than the ones related to the construct being tested, and the difficulty factor adversely affects examinees' performance on the test (Camilli & Shepard, 1994).

However, the differences in item performance are not evidence of item bias. Group difference in item performance can represent exact knowledge and experience differences with respect to the purpose of the test (Gierl et al., 1999; Huang & Sheeran, 2011). This outcome is referred to as *item impact*. Like bias, impact is constant for the members of a particular group, but these effects reflect performance differences that the test is intended to measure (Clauser & Mazor, 1998).

3. Types of DIF

There are two different types of DIF: uniform and non-uniform DIF. Mellenbergh (1994) distinguished both uniform and non-uniform DIF. Uniform DIF occurs when there is no interaction between ability level (θ) and group membership. That is, the probability of answering the item correctly is uniformly greater for one group than another group over all levels of ability. That is to say, for uniform DIF items only the difficulty parameter (b_i) is different between groups but the discrimination parameter (a_i) is the same. Non-uniform DIF occurs when there is interaction between ability level (θ) and group membership. That is, the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels. Non-uniform DIF is reflected by item characteristics curves (ICCs) that are not parallel. Therefore, the discrimination parameter is different between groups but the difficulty parameter can be the same or different.

Non-uniform DIF can be further divided into two kinds of interaction: Disordinal and ordinal. Disordinal interaction between ability and group membership is indicated by ICCs that cross in the middle of the ability range. Alternatively, ordinal interaction is indicated by ICCs that cross at either the low end or the high end of the ability scale, resulting in ICCs that appear to be similar over most part of the ability range (Dickinson et al., 2003).

4. DIF Analysis Procedures

The DIF analysis produces statistics describing the amount of DIF for each test item. The analysis also produces statistics describing the statistical significance of the DIF effect – the probability of finding so large an effect in the available samples of examinees if there were no DIF in the population from which they were sampled. The decision rule based on the

Mantel-Haenszel procedure sorts out the test items into three categories, labeled A (least DIF), B (moderate), and C (most DIF) (Holland & Thayer, 1988).

All items whose statistics place as "high B" or "C" are reviewed by content specialists and test developers for the purpose of identifying item bias. Content reviewers try to examine each "high B" and "C" item and try to determine whether its DIF can be explained by characteristics of the item that are unrelated to the measurement purpose of the test. If it can, the item is deleted from the scoring of the test (Camilli & Shepard, 1994).

It is important to realize that DIF by itself is not considered sufficient grounds for removing an item from the exam. The item may test an important piece of knowledge that happens to be more common in one group than another. Only if the DIF is attributable to factors other than the knowledge being tested is it grounds for deleting the item (Gierl et al., 1999; Huang & Sheeran, 2011).

5. Different DIF Detecting Procedures

Several different methods are currently used to determine whether a test item displays DIF. As many of those methods are becoming more popular, none has been generally accepted by the assessment professionals (Clauser & Mazor, 1998; Holland & Thayer, 1988).

Many item response theory (IRT) based methods have been proposed to identify DIF. Of those, the method suggested by Raju, van der Linden, and Fleer (1995) and the one studied by Thissen, Steinberg, and Wainer (1993) are the popular ones. Those methods require large sample size especially when dealing with a three-parameter model (Clauser & Mazor, 1998).

The non-compensatory DIF (NCDIF) is a purely item level statistic that reflects true score differences for the two groups under examination. The NCDIF index considers each item separately, without considering the functioning of other items in the scale. In mathematical terms, NCDIF is the square of the difference in true scores for the two groups, averaged across the ability level. Thus, the square root of NCDIF gives the expected difference in item responses for individuals with the same standing on the ability scale, but belonging to different groups. This index tends to identify items for which the area between the two ICCs, one for the reference group and one for the focal group, is excessively high. Since the chi-square statistic is influenced by sample size, the box and whiskers plot of the NCDIF values allow to categorize DIF items in relation to their relative severity (Raju et al., 1995).

Model comparison approach (Thissen et al., 1993) consists in testing DIF for each item by estimating the n item parameters. The statistic used is called LOGDIF. The procedure specifies that an item presents DIF when the value of LOGDIF leads to a statistical significance. Since the chi-square statistic is influenced by sample size, the box and whiskers plot of the LOGDIF values allow to categorize DIF items in relation to their relative severity.

During the last decades, several non-IRT based DIF methods were proposed. Of those methods, three seem particularly promising: the Mantel-Haenszel method (Holland & Thayer, 1988), the Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993), and the logistic

regression method (Clauser & Mazor, 1998).

Holland & Thayer (1988) proposed a statistic, previously discussed by Mantel & Haenszel (1959), to develop a method for detecting DIF. Throughout the years, this method became more and more popular (Zwick, 1997). The Mantel-Haenszel method compares, for a given item, the probability of obtaining a right answer in the focal group to the probability of obtaining a right answer in the reference group for subjects of equal ability. There are many ways to determine the presence of DIF using the Mantel-Haenszel method (Roussos, Schnipke, & Pashley, 1999). For example, the most commonly used method is to calculate the value of D_{MH} = -2.35 ln(a_{MH}). If the absolute value D_{MH} is higher than 1.5 and significantly higher than 1 at the 0.05 level of significance, the item is classified as category C (most DIF). If the absolute value of D_{MH} is lower than 1 and not significantly higher than 0 (at a =.05), the item is classified as category A (least DIF). For any other situation, the item is classified as category B (moderate).

The SIBTEST is a non-parametric procedure. It provides an effect size measure and a test of significance. The effect size measure, $\hat{\beta}_U$ is estimated by $\hat{\beta}_U = \sum_{K=0}^{n} \hat{P}_k (\overline{Y}^*_{R_k} - \overline{Y}^*_{F_k})$, where $\hat{P}_k$ is the proportion of focal examinees at each score point *k*, and $\overline{Y}^*_{R_k}$ is the estimated true score for the reference group and $\overline{Y}^*_{F_k}$ is the estimated true score of the focal group at each score point *k*. The estimated true scores are produced using a regression correction described by Shealy and Stout (1993). If the estimated effect size, $\hat{\beta}_U$ is positive, then the items favors the reference group. In contrast, if it is negative, then the item favors the focal group.

Roussos and Stout (1996) suggested a range of values for interpreting the amount of DIF: 1) least or A-level DIF: absolute value of $\hat{\beta}_U$ < 0.059 and H_0: B = 0 is rejected, 2) moderate or B-level DIF: absolute value of 0.059 < $\hat{\beta}_U$ < 0.088 and H_0: B = 0 is rejected, and 3) most or C-level DIF: absolute value of $\hat{\beta}_U \geq$ 0.088 and H_0: B = 0 is rejected. These guidelines were used in the current study to identify DIF items.

Logistic regression (Swaminathan & Rogers, 1990) allowed the development of a DIF detection method that is gaining in popularity (Clauser & Mazor, 1998). This method is commonly used to identify DIF (Gierl et al., 1999). The logistic regression model consists of two stages. First, the control variable, usually the "classical" total score, is to be included in the regression equation. Then two other variables, related to the group and the interaction group score, are included in the equation. The analysis consists in testing if the insertion of these two variables leads to a significant statistical result. If it is positive, then DIF is present.

This method is based on chi-square statistics. Zumbo and Thomas (1996) created an index for DIF level classification for logistic regression approach based on partitioning a weighted least squares estimate of R^2 that yields an effect size measure. Jodoin (1999) proposed guidelines for interpreting $R^2\Delta$. If the chi-square value for a given item is statistically significant or when $R^2\Delta$ <0.035, it will be said that the item has A-level DIF. If the chi-square value is statistically significant and when 0.035< $R^2\Delta$ <0.070, it will be said that the item presents B-level DIF. If the chi-square value is statistically significant and when $R^2\Delta$ >0.070, it will be said that the item has C-level DIF (Gierl et al., 1999).

Currently, all those methods are popular but few studies were implemented to compare the stability of these procedures. Many of the research done on the topic mainly focused on the stability of non-IRT based methods (Huang, 1998; Ackerman & Evans, 1992). Though a number of studies indicated that IRT-based methods are indeed capable to detect DIF (Clauser & Mazor, 1998), it has yet to be researched whether IRT based methods are more stable then non IRT based methods.

6. Causes of DIF

Current DIF research has employed different methods to detect DIF across genders (Dodeen & Johanson, 2001; Henderson, 2001; Kranzler & Miller, 1999), races (Sammons, 1995; Zhang, 2002), and languages (Gierl & Khaliq, 2001; Hamilton, 1999; Huang & Sheeran, 2011; Walstad & Robson, 1997). But very few of the studies have included substantial content review procedures, though a couple of translation DIF studies have identified some sources of translation DIF. For example, some researchers used substantive analyses in their studies and identified several sources of translation DIF: Omissions or additions that affect meaning; differences in words or expressions inherent to language or culture; differences in words or expressions not inherent to language and culture; and format differences (Alalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001; Huang & Sheeran, 2011).

Bias may occur, for example, when the meaning of an item changes after the test is translated. Hambleton (1994, p. 235) provided an example about this problem. In a Swedish-English comparison, English-speaking examinees were presented with the following item:

Where is a bird with webbed feet most likely to live?

A. in the mountains

B. in the woods

C. in the area

D. in the desert

In the Swedish test, the phrase “webbed feet” was translated to “swimming feet,” thus providing an obvious clue to the Swedish-speaking examinees about the correct option for this item.

Translation may also cause bias if place names are used inconsistently across the source and target languages. Huang & Sheeran (2011) provided the following example. In an English-Chinese comparison, Chinese-speaking examinees were presented with the following item:

*在高山地區如西藏高原**, 可以開發的再生能源不包括*

A 地熱能 *B 潮汐能**

C 風力能 *D 太陽能*

whereas the English-speaking examinees were presented with the following item; *In a high mountain area such as **Xizang Gaoyuan*** in China, the following types of renewable energy may be generated except*

A geothermal energy *B tidal energy**

C wind energy *D solar energy*

where * indicates the right answer and ** indicates the unequivalent part.

In the above example, the inconsistent use of place names across the two languages (English and Chinese) becomes an important source of translation DIF for this particular content area of geography. For example, both *pinyin* (the sound of Chinese characters with no association with meaning) and actual characters (having exact meaning) about place names were used inconsistently across the two language groups. The original English question asked about "In a high mountain area such as Xizang Gaoyuan in China, the following types of renewable energy may be generated except…" and the right answer was "tidal energy." The original English version used the word "Gaoyuan" (Chinese *pinyin* for the word "plateau"), whereas the translated Chinese version used the word "plateau." As a result, this item became "a biased item which produced the highest absolute value of $\hat{\beta}_U$ (= 0.349). The highest Beta estimate indicates that this item produced the largest performance difference between the Chinese- and English-speaking examinees. Statistical results show that the difficulty level of this item was -0.991 for the Chinese examinees but 0.582 for the English examinees, and about 80.1% of the Chinese examinees got this question correct but only 48.1% of the English examinees answered this question correct" (Huang & Sheeran, 2011, pp. 26-27).

Substantive research focused on why DIF occurs while little success has been made. "Attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure" (Roussos & Stout, 1996, p. 360). That is to say, although DIF procedures may help improve test quality and increase test fairness, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. It is very hard to find out the reasons for the differential performance on test items or to identify a common deficiency among the identified items.

The potential substantive reasons for DIF are still largely unknown, according to Roussos & Stout (1996). Related literature suggests some common and most widely

discussed explanation for the occurrence of DIF: examinees' familiarity with the content of the items, variously referred to as exposure, experience, or cultural loading (Eells, Davis, Havighurst, Herrick, & Tyler, 1951; Jensen, 1980; Reynolds & Kaiser, 1990). Other explanations include: the examinees' interest in the content (Eells et al., 1951), and their negative emotional reaction, for example, anger, disgust, fear, to the content (Wendler & Carlton, 1987). All these explanations have been offered to account for items with DIF in research studies (Donlon, Hicks, & Wallmark, 1980; Harris & Carlton, 1993) as well as in test development procedures that evaluate the suitability of specific items for operational use (Zieky, 1993). However, little or no empirical evidence is available to support these conjectures about group difference by gender (between males and females), by race (between White and Black examinees), by first language (between English and French), and so on in their reactions to test items. Establishing the existence of such group differences and their connections with DIF would not only contribute to our understanding of this phenomenon but would also guide further research into what causes DIF and what can be done to minimize it (Stricker & Emmerich, 1999).

7. Critique of Current DIF Research

As previously discussed, there are a number of procedures for detecting DIF. However, these methods cannot produce consistent and stable results. For example, in a study by Gierl et al. (1999), the researchers used the Mantel-Haenszel method, SIBTEST, and logistic regression procedure, but all the three procedures produced inconsistent results. Therefore, current DIF procedures cannot provide accurate and reliable statistical results and further research needs to be conducted in the area of exploring the most effective or accurate DIF detecting statistical procedures.

In addition, most of the current DIF research deals with only the detection of DIF items. According to the model proposed by Camilli and Shepard (1994), judgmental analysis needs to be done in order to find out the causes of DIF. Test items can then be modified and fairer tests constructed. That is to say, a majority of the DIF studies are comparing and contrasting different DIF procedures. There is a clear lack of research component on why DIF occurs. Further, the few studies which tried to explain why DIF occurred indicated that it was hard for the content reviewers to identify the exact causes for some DIF items. It is believed that there is not a strong theoretical framework for the explanation of DIF. In order to better explain why DIF occurs, cognition needs to be studied. Cognitive explanation may provide us with answers to identified DIF problems.

Further, the current content review procedures involve only test developers, content specialists, and linguists. It is suggested that sometimes both the teachers and examinees be invited to join the content review committee. They may come up with better explanation for specific test items which display DIF.

Finally, the impact of DIF should be appropriately interpreted. It cannot be overestimated; and it cannot be underestimated, either. DIF is not always a bad thing. Only if a DIF item is a biased item, it can then be rewritten or completely removed from the test. If the DIF item is only an item impact, this is exactly what the item is intended to measure

(Clauser & Mazor, 1998).

8. New Directions for DIF Research

It is suggested that future DIF research be strengthened in the following four areas: a) to find more accurate detecting procedures; b) to build stronger theoretical frameworks; c) to explore more effective content review methods; and d) to establish more appropriate interpreting guidelines.

To Find More Accurate Detecting Procedures

There are already many statistical methods for detecting DIF. These methods have different strengths and weaknesses. For example, some methods can only be used to detect uniform DIF, but not non-uniform DIF. Some can only work with dichotomous data, but not polytomous data. Some methods are more powerful in detecting DIF than other methods. Future research can focus on the selection of DIF detecting methods, namely, how to choose the most appropriate statistical methods according to specific research contexts.

Different methods provide different results, which might cause concerns for the consistency of results. In order to produce consistent research results, there should be some specifications for the selection of statistical methods. Further, it is important that the corresponding computer software should be available, and easy to access and use.

To Build Stronger Theoretical Frameworks

One purpose of DIF research is to understand it. In other words, why does it occur? There is no strong theoretical framework to support current DIF research. The lack of theoretical framework limits our interpretation of the causes of DIF. It is suggested that the research on the theoretical framework for DIF research be interdisciplinary in nature. For example, issues associated with DIF should be considered and studied from multiple perspectives: cognitive, cultural, linguistic, social, pedagogical, and historical. In other words, the occurrence of DIF is due to multiple sources. The ignorance of any source leads to the failure of building a sound theoretical framework.

To Explore More Effective Content Review Methods

Content review is an important step to separate item biasness from item impact. Using a strong DIF theoretical framework, the researchers can conduct more thorough content review. The content review process should involve different professionals with diverse expertise. Not only content experts and test developers are involved, educational psychologists, instructors, and even student representatives should be involved. This is because different people can provide different explanations from different perspectives.

To Establish More Appropriate Interpreting Guidelines

The impact of DIF cannot be overestimated or underestimated. How should DIF items then be appropriately interpreted? First, it is important to remember that a DIF item does not necessarily suggest biasness. Only through specific procedures can it be determined whether it is item biasness or item impact. It is very natural that some test items display DIF. What

needs to be done is to figure out the causes of DIF. A correct decision could then be made about the DIF item: either to remove it from the test if it is biased or keep it in the test if it is not biased.

References

Ackerman, T. A., & Evans, J. A. (1992). *An investigation of the relationship between reliability, power, and Type I error rate of the Mantel-Haenszel and the simultaneous item bias detection procedures*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Alalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*, 185-198. http://dx.doi.org/10.1111/j.1745-3984.1999.tb00553.x

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test item*. Newbury Park, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44. http://dx.doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01132.x

Dickinson, T. L., Wanichtanom, R., & Coates, G. D. (2003, April). Differential item functioning: Item response theory and confirmatory factor analysis. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, April 11-13, Orlando, Florida.

Dodeen, H., & Johanson, G. (2001). *The prevalence of gender DIF in survey data.* Paper presented at the Annual meeting of the American Educational Research Association, Seattle, WA.

Donlon, T. F., Hicks, M. M., & Wallmark, M. M. (1980). Sex differences in item response on the Graduate Record Exam. *Applied Psychological Measurement, 4*, 9-20. http://dx.doi.org/10.1177/014662168000400103

Eells, K., Davis, A., Havighurst, R. J., Herrick, V.E., & Tyler, R. W. (1951*). Intelligence and cultural differences.* Chicago: University of Chicago Press.

Gierl, M. J., & Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1(3&4)*, 249-270. http://dx.doi.org/10.1080/15305058.2001.9669474

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

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *The Alberta Journal of Educational Research, 14*, 353-376.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: Aprogress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hamilton, L. S. (1999). Detecting gender-based Differential Item Functioning on a constructed-response science test. *Applied Measurement in Education, 36*, 1-28.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*, 137-151. http://dx.doi.org/10.1207/s15324818ame0602_3

Henderson, D. L. (2001). *Prevalence of gender DIF in mixed format high school exit examinations*. Paper presented at the Annual Meeting of the American Educational Research Association. Seattle, WA.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, C. Y. (1998). *Factors influencing the reliability of dif detection methods*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Huang, J., & Sheeran, T. (2011). Identifying causes of English-Chinese translation differential item functioning. *International Journal of Applied Educational Studies*, *12*(1), 16-32.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Macmillan Publishing, Co.

Jodoin, M. (1999). *Reducing type I error rates using an effect size measure with the logistic regression DIF procedure*. Unpublished master's thesis, University of Alberta. Edmonton, Alberta.

Kranzler, J. H., & Miller, M. D. (1999). *An examination of racial/ethnic and gender bias on curriculum-based measurement of reading*. Research/technical report in Reading and communication skills clearinghouse, ERIC accession number: ED435087.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741-746. http://dx.doi.org/10.2307/3587128

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NY: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 30. http://dx.doi.org/10.1037/0033-2909.115.2.300

Raju, N. S, van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 4, 353-368. http://dx.doi.org/10.1177/014662169501900405

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 487-525). New York: Wiley.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230. http://dx.doi.org/10.1111/j.1745-3984.1996.tb00490.x

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 3, 293-322. http://dx.doi.org/10.2307/1165326

Sammons, P. (1995). Gender, ethnic and socio-economic differences in attainment and progress: A longitudinal analysis of student achievement over 9 years. *British Educational Research Journal, 21*, 465-485. http://dx.doi.org/10.1080/0141192950210403

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias / DIF as well as item bias / DIF. *Psychometrika, 58*, 159-194. http://dx.doi.org/10.1007/BF02294572

Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement, 4*, 347-366. http://dx.doi.org/10.1111/j.1745-3984.1999.tb00561.x

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using the logistic regression procedure. *Journal of Educational Measurement, 27*, 361-370. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in Economics. *Journal of Economic Education, 28*, 155-172. http://dx.doi.org/10.1080/00220489709595917

Wendler, C. L. W., & Carlton, S. R. (1987, April). *An examination of SAT-verbal items for differential performance by women and men: An exploratory study.* Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Zhang, Y. (2002). *DIF in a large scale mathematics assessment: The interaction of gender and ethnicity*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Ziecky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland and H. Wainer (Eds.), *Differential Item Functioning: Theory and Practice*. Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D., & Thomas, D. R. (1996). *A measure of DIF effect size using logistic regression procedures.* Paper presented at the National Board of Medical Examiners, Philadelphia, PA.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and Psychological Measurement, 57*, 3, 412-421. http://dx.doi.org/10.1177/0013164497057003003

Copyright Disclaimer