Differential Item Functioning: Implications for English Testing in China

Ligang Liu[1] and Song Wang[1,*]

[1] *Teaching Department of Foreign Languages, Chengdu Medical College, Chengdu, Sichuan 610500, China*
**Corresponding author. Email: wsong80@126.com*

ABSTRACT
This paper aims to recapitulate an important concept in test validity and fairness, namely Differential Item Functioning (DIF). Then comes the discussion about the present DIF studies in the testing of English as a foreign language. Finally, after pointing out the limits and problems of the present DIF study on English testing in China, suggestions are given in this paper which indicates that the DIF study on English proficiency tests in China should focus on different education conditions, different academic backgrounds, and different DIF levels.
Keywords*: Differential Item Functioning, test validity, test fairness, English testing*

1. INTRODUCTION

Language use is a very complicated process and thus cannot be treated as a fully uniform repertoire. Language proficiency is not the only factor deciding or influencing the performance of a language user. In fact, a number of factors have impact on language use or performance, including various kinds of contexts of use and the language user's own characteristics like social and geographical origin, age, gender, education, occupation and so on. It is these factors that cause difficulties in the development of language ability tests designed to measure language users' language ability. English language testing, as a foreign language proficiency test, is widely used in China to measure the overall English language ability of Chinese English learners. Since the inference or interpretation of the scores of these tests are vital in decisions about admission, promotion, degree awarding etc., test quality and test fairness are of great importance to ensure all test takers have equal opportunities. According to Kunnan (2004), test fairness comprises validity, access, justice etc. Among these factors concerning test fairness, validity is an essential feature of tests' interpretation and use, which involves construct, content, criterion, consequential and face validity and so on. To improve test validity, Differential Item Functioning (DIF) has attracted much attention due to its ability to ensure the metric equivalence of measurement instruments. DIF works on the principle that different groups of test takers (for instance, males vs. females, native vs. non-native speakers, or students of sciences vs. students of humanities) with a similar level of knowledge should perform similarly on individual test items regardless of group membership.

2. DIFFERENTIAL ITEM FUNCTIONING (DIF)

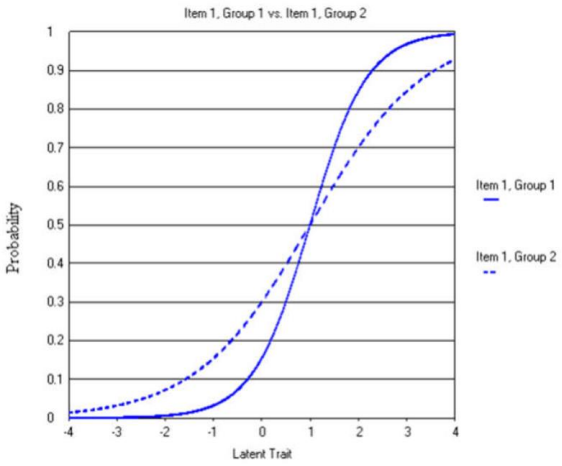
2.1. Concept of DIF

In the field of language assessment, the term "fairness" has 2 connotations: one is measurement of fairness, which concerns the quality and validity of a test per se; the other is the concept of social fairness, which is fundamentally a post-test social issue imposing a certain responsibility on the test users to justify how the test results are actually used. Since fairness is a necessary and important perspective of test validity, a test is unfair or not free from bias if one group is systematically disadvantaged over another, not because of their knowledge or ability, but because of other factors such as item construct, item content, gender, native language, major backgrounds and so on. Differential Item Functioning (DIF), as an item-level performance difference between groups of examinees or applicants matched on a latent trait, has received various definitions or discussions in the field of psychometric measurement (Salehi & Tayebi 2012). In language testing, DIF means that examinees with similar language knowledge or proficiency may be given different results because some or many test items advantage one group or disadvantage another. Thus test items are biased since they cause difficulty to particular examinees which is irrelevant to what is being measured and leads to a discrimination against these examinees. Actually, DIF has been widely used in test validation process and become an integral part of test fairness. A test item will be considered to exhibit DIF when it works differentially for a particular group of examinees, depriving them of their equal chance of a successful response to an item (Zumbo, 2003).
Originally, the term of DIF refers to the performance difference in a single item between groups of examinees

matched on a latent trait. With the progress made in DIF studies, the notion of DIF is expanded from item level to bundle or test level, i.e., Differential Bundle Functioning (DBF) or Differential Test Functioning (DTF). DBF is a notion built on DIF, in which a subset of items in a test are organized to form a group of two or more items and this group of two or more items are analyzed for performance difference among test takers of different membership, after controlling their overall ability (Latifi et al., 2016). There should be a specific organizing standard or principle for researchers to organize the test items to form a bundle. For instance, test items can be grouped based on their shared construct, common content, or similar expression etc. Unlike the item-level notion of DIF, DTF is a test-level notion which refers to the fact that test bias occurs when the expected true score at the scale level is not the same for two groups of examinees (Drasgow, 1987) or when measurement equivalence of a test does not hold for two groups f examinees (Zumbo, 2003). DTF is a very important notion in the study of test validity and test fairness because interpretations of and decisions with a test are made not according to examinees' performance on an individual item but according to examinees' performance on a whole test. It is still unknown whether DIF can cause cumulative effect on DTF and what the relationship is between DIF and DTF.

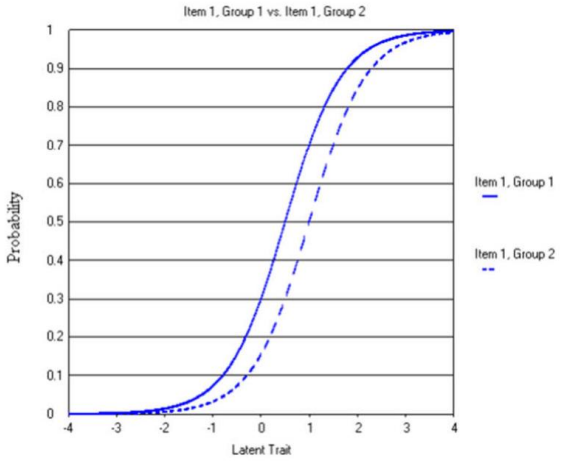
2.2. Different Types of DIF

Generally speaking, DIF is divided into two types, uniform and non-uniform DIF. Uniform DIF is also called unidirectional DIF, which occurs when the probability of a correct response for one group is greater than for the other group over all the levels of proficiency (Ibrahim 2018). Non-uniform DIF is also called crossing DIF, which exists when the difference in the probability of a correct response is not the same at all levels of proficiency between the two comparison groups (Ibrahim 2018). That is to say, in non-uniform DIF case, the probability of correctly answering an item is higher for one group at a certain level of proficiency, while it is higher for the other group at other levels of proficiency. The difference between uniform and non-uniform DIF can be showed in figure 1 and figure 2. In these two figures, if the item response curves of the two comparison groups intersect with each other, non-uniform DIF occurs (figure 1); if the item response curves of the two comparison groups do not cross, uniform DIF occurs (figure 2).



Source: Zwick, Donoghue and Grima (1993)

Figure 1. An item showing non-uniform DIF



Source: Zwick, Donoghue and Grima (1993)

Figure 2. An item showing uniform DIF

In figure 1 where non-uniform DIF exists, the full curve indicates group 1 and the dotted line indicates group 2. The space between the 2 curves in figure 1 indicates the existence of non-uniform DIF. Below the point where the 2 curves cross, the space suggests that the item is easier for group 2 than group 1 at lower ability levels; above the point where the 2 curves cross, the space suggests that the item is harder for group 2 than group 1 at higher ability levels. However, in figure 2 where uniform DIF exists, the space between the 2 curves indicates the existence of uniform DIF. The space suggests that the item is harder for group 2 than group 1 at all ability levels.

2.3. Methods of DIF Detection

The DIF analysis framework is based on the concept of primary and secondary dimensions. Dimension here refers to a fundamental characteristic of a test item that can play a

part in the odds of getting the correct response. It has been found in past studies that test scores are often affected by a secondary dimension besides the primary dimension being measured. Every item in a test is designed to measure the main construct seen as the primary dimension. When DIF occurs, the item showing DIF measures at least one secondary dimension in addition to the primary dimension (Shealy & Stout, 1993). This kind of secondary dimensions may be either auxiliary or nuisance. A secondary dimension is auxiliary if it is intended to be operationalized (e.g., vocabulary knowledge in a reading test); a secondary dimension is nuisance if it is intrusive to learners' test performance and contaminative to their scores (e.g., reading proficiency in a mathematics test). DIF resulting from auxiliary dimensions is benign, while DIF resulting from nuisance dimensions is adverse and thus reflects bias. DIF detection is typically based on the comparison of 2 groups, i.e. the reference group and the focal group (these 2 groups differ in their group membership such as gender, education, academic major, age etc.). Bias can be detected in items by comparing the focal group and the reference group.

Albeit there are various kinds of statistical methods available for DIF detection, just like there is no single key for all the doors, there is no single DIF method effective or ideal for all study purposes since different DIF methods flag DIF differently and have different requirements for sample size and time consuming.

Classical DIF detection methods are the Mantel–Haenszel procedure (MH) (Holland & Thayer, 1988) and the Standardization method (Dorans & Kulick, 1986). There are also a few approaches developed based on Item Response Theory (IRT) such as Logistic Regression (Log-R) (Swaminathan & Rogers, 1990), Area Measures (Raju, 1988) , Wald Statistics (Lord, 1980) and Loglinear Item Response Models (Mellenbergh, 1982). However, quite a few IRT–based approaches are sensitive to sample size and model-data fit and time–consuming. There are other methods which are often used in DIF detection like the Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993) and the Rasch Model (RM) (Bond & Fox 2015). Among all these DIF detection methods, only a relatively small number of them are preferred and widely used because of their theoretical and practical strengths. The MH, the Log-R, the SIBTEST and the RM are the most popular DIF detection methods. The MH procedure is popular due to the fact that it is less sensitive to sample size and can be used to analyze both dichotomous (binary type, right or wrong) and polytomous (ordinal type, grade 1, 2, 3, 4, 5) data. The Log-R is popular based on the fact that it is very useful in detecting both uniform and non-uniform DIF, and can also be used for dichotomous and polytomous data while requiring less complicated computing than other IRT-based approaches. The advantage of the RM technique is that it can be applied to both large and small samples, and it even shows higher DIF detection rates than the MH when the sample size is small. The SIBTEST technique is popular for the fact that it uses original item response data rather than parametric estimates required for IRT models and it is more effective in DIF detection than the MH and the Log-R.

3. PRESENT STUDIES ON DIF

A lot of DIF studies have been conducted among test takers from different linguistic, cultural, ethnic, racial backgrounds to improve test quality and test fairness in international language tests, especially those high stakes tests. The most popular studies focus on whether DIF occurs among test takers from different native language backgrounds (Kim 2001; Uiterwijk &Vallen 2005). This is because it is suggested by many findings that the major influence in second language acquisition is language learners' native language. So, a language test for native speakers may be more difficult and challenging for second language learners, and a language test for second language learners (e.g. IELTS) may be less challenging for some second language learners (e.g. German or French speaking English learners) and more challenging for others (e.g. Arabic or Chinese speaking English learners).

Test takers' characteristics are also those major factors related to test bias in DIF study. Many studies have focused on whether DIF occurs among test takers from different gender groups (Ryan & Bachman 1992), from different age groups (Geranpayeh & Kunnan 2007), or from different academic background groups (Pae 2004). DIF caused by gender difference or academic background difference may result from the fact that males or students of sciences are probably more familiar with the topic of reading or listening materials in the test since males' preference for popular movies and readings are different from females and students of sciences major in different subjects from students of humanities. DIF caused by age difference may come from the fact that test takers of different age have different language educational experiences or different language skills like reading or listening. Therefore, certain language skills also become the focus of test bias study. For example, Li and Suen (2012) developed DIF into DSF (Differential Skill Functioning) and focused on whether test takers with different native language background are better at certain language skills. It is found that different language skills (e.g. vocabulary, syntax, extracting explicit information, connecting and synthesizing) may favor test takers of different gender or test takers of different native language background, which can provide important guidance for language instruction and learning.

4. DIF STUDIES IN CHINA

China has the largest population in the world, and also has the largest number of learners of English as a foreign language. In China, there are various kinds of English examinations, including those developed independently by China, such as English Test of High School Entrance Examination, English Test of College Entrance

Examination, English Test of Postgraduate Entrance Examination, College English Test: Band 4 (CET4), College English Test: Band 6 (CET6), Test for English Majors: Band 4 (TEM 4), Test for English Majors: Band 8 (TEM 8). The tests given above are all high stakes tests, the score use and interpretation of which plays a vital part in decisions about admission, employment, graduation, etc. Thus it is very important to improve the validity and fairness of these tests. However, China is a large country with a large population, and there is disparity in economic and educational development between different areas, especially between urban and rural area. Test takers in China differ not only in gender and academic background, but also in educational background and English learning conditions and experiences. Therefore, for those English tests in China, especially those national unified English tests, DIF study is necessary and vital in validation process and fairness improvement.

As DIF has been receiving more and more attention in the study of test validity and fairness, DIF is also used in developing standardized tests in China. Li and Kong (2009) made a study on bias in districts and gender in 2005 TEM 4 reading comprehension section, since there is a big imbalance between male and female test takers in this test, and a big difference in economy and education between eastern and western areas in China. Lei (2007) made an analysis on the difference between males and females in the English achievements of Shanghai College Entrance Examination, which found that the achievement difference did not come from DIF, but from the difference between males and females in the development of English language proficiency. Dong and Ma (2001) made a comparison between 3 usual DIF detecting methods based on 75 items of multiple choice in 1999 English test of College Entrance Examination. So far, the DIF studies on English tests in China have focused mainly on gender or district difference, so the perspectives these studies focus on are rather limited, and more DIF studies should focus on ethnic, education, and academic background difference to detect whether there are construct-irrelevant factors or the construct being measured is underrepresented.

5. SUGGESTIONS FOR DIF STUDIES IN CHINA

Concerning the English proficiency tests in China, more DIF studies are needed to enhance test validity and test fairness. Such studies should focus on different education levels and different academic backgrounds. Simultaneously, studies on test bias (DTF: Differential Test Functioning) and its relationship to DIF are also of vital importance in test validity and test fairness.

5.1. Importance should Be Attached to Education Difference in DIF Study

As stated above, China is a large country with an imbalanced economic development which also results in an imbalanced educational development. Firstly, there is disparity in educational development and education conditions between eastern and western areas. Eastern areas, centered on Shanghai, have advantage in economy over western areas since economic development in western areas is hindered by their natural conditions, such as mountains or deserts. So compared to the eastern areas, the education investment in western areas is far from enough, which results in insufficient teaching staff and poor education conditions. Secondly, there is disparity in education conditions between urban and rural areas. Similarly, the disparity in economic development between urban and rural areas leads to the disparity in educational development between urban and rural areas. And urban areas are more attractive to young and talented teachers because of the fast development of urbanization. Thus young and talented teachers are only a small part of the whole teaching staff in rural areas, and it is not easy for the teachers in rural areas to change their teaching ideas and methods. Thirdly, there is still disparity in education conditions between major and minor cities. Today in China, major cities such as Shanghai, Shenzhen, Guangzhou, Nanjing, and Chongqing, are also economic centers in their respective local areas. As economic center in their local region, major cities often have advantages over their nearby minor cities in employment, education investment, health care and so on. To be a teacher in major cities will provide one person both a higher salary and a better future in his/her teaching career.

Different education level or condition should be another focus in DIF studies in English testing in China. As for those national unified English tests like CET 4, CET 6, TEM 4 and TEM 8, the test takers are from various education conditions because of the disparity in educational development between eastern and western areas, between urban and rural areas, and between major and minor cities. Specifically speaking, English is a foreign language in China, and foreign language learning is different from native language learning because for most foreign language learners, there is a lack of opportunities to practice foreign language in everyday communication. However, Chinese English learners in urban areas and major cities have easier and better access to talented English teachers, teachers of native speakers, and latest reading or listening materials. So for English education, there must be a disparity between urban and rural areas, and between major and minor cities. Thus a series of DIF studies are supposed to be done on test takers from areas of different education conditions. It is first suggested that DIF studies be conducted in oral English test and listening tests concerning test takers between urban and rural areas, or between major and minor cities, because it is usually thought that English learners in urban areas and major

cities have more opportunities to practice English speaking and listening.

5.2. More Attention should Be Paid to Academic Background Difference in DIF Study

There have been only very few DIF studies focusing on different academic background of test takers, and such DIF studies have not been seen in China. In China, high school students are divided into students of sciences and students of humanities. However, when they take College Entrance Examination for higher education, they will take the same English subtest of College Entrance Examination. Is there any item or are there any items in the English subtest of College Entrance Examination which function differentially between students of sciences and students of humanities? There has been no answer for this question. This is also the case for the English subtest of Postgraduate Entrance Examination, College English Test: Band 4 (CET4), and College English Test: Band 6 (CET6). Because all the college students in China (except English majors: English majors take TEM 4: Test for English Majors Band 4 and TEM 8: Test for English Majors Band 8) take CET4 and CET6, there is great necessity for DIF studies of CET4 and CET6 to concentrate on different academic background. It is suggested in this paper that DIF studies of CET4 and CET6 should concentrate not only on the difference between sciences and humanities, but on the difference between engineering, science, medicine, agronomy, economics, literature etc. This is because there are various kinds of professional disciplines in colleges and universities in China, and the science-humanity division in high school is relatively rough and may not be sensitive enough in DIF detecting studies focusing on academic background difference in college English test takers. Therefore, in future DIF studies of college English testing in China like CET4 and CET6, DIF researchers are supposed to have a more specific division of test takers' academic backgrounds to lessen the test items' possible impact on the performance of test takers of different academic backgrounds.

5.3. DTF Studies should Be Given More Concern

Since DBF refers to the group difference in the results of score on a subset of test items after the controlling for the ability, there is actually no precise boundary for DIF and DBF study because a lot of DIF studies usually take many items into scrutiny. For example, the reading section under scrutiny in Li and Kong (2009) is made up of 4 passages and altogether 20 items, which aim to measure the examinees' ability to discern facts and details, the ability to make an inference about the author's implied intention, the ability to summarize the main idea of the article, and the ability to infer the meaning of the words according to the context. Researchers can group into a bundle those items measuring the same ability construct. Dong and Ma (2001) put 75 items under scrutiny. All these 75 items are of dichotomous type and in the form of multiple choice. Although these 75 items include different sections such as vocabulary, grammar, and reading, the researchers treat the all 75 items as one bundle. The methods used for DIF detection such as SIBTEST (Simultaneous Item Bias Test) can also be used for DBF detection. One advantage of DBF study is that DBF effect can be significant when the DIF effect of every single item is minimal, or DBF effect can be minimal when the DIF effect of every single item is significant.

There is a lack of studies on DTF and the relationship between DIF and DTF is still unclear. Although there has been an increasing number of DIF studies on reading, vocabulary and grammar section of English tests in China, almost no study has examined the effect of DIF on test level bias. However, the present DTF studies have produced mixed or even contradictory results. Some studies identify several items as showing DIF in different sections such as listening and reading, while at the test level these studies identify negligible effects, which indicates a possible DIF cancellation effect (Pae 2004; Zumbo 2003). Pae and Park' study (2006) indicates that the effect of item level DIF, once detected, can be carried to the scale or test level bias regardless of the DIF directions. Pae and Park (2006) uses more advanced analytical techniques, which may be the cause for the different results of the DTF studies. But this also suggests the relationship between DIF and DTF may be more complicated than what researchers used to think. In this regard, more concern should be given to the study on DTF and its relationship with DIF. Particularly in China, English tests have not seen such studies on DTF and its relationship with DIF. Although the present DIF studies (Li & Kong 2009; Dong & Ma 2001) have not identified significant DIF effect on item level in some sections like reading or listening in English tests in China, it is unknown whether the minimal DIF effect on item level may carry a cumulative effect on test level in such English tests due to a lack of DTF studies in China. Therefore, not only are different types of DIF study (on different education conditions, academic backgrounds etc.) in great need, DTF studies are also in urgent need to ensure the validity and fairness of English tests in China. For future DTF studies in China, DTF or test bias can be investigated from 3 perspectives. Firstly, it can be investigated by studying the association between the test score and an external criterion (Jensen, 1980); secondly, by computing expected true scores for two groups of examinees using the IRT Test Characteristic Curve (TCC) (Drasgow, 1987); thirdly, by comparing internal factor structures across identifiable subgroups of examinees (Zumbo, 2003).

6. CONCLUSION

Differential Item Functioning (DIF), as a very important notion in psychometric measurement, is an essential part in validation process and test fairness. English testing in China has already seen an increasing number of DIF studies, but the perspectives from which these studies are conducted are still limited. DIF studies concerning English testing in China should also focus on examinees' different education conditions and academic backgrounds. Simultaneously, DIF should not only be studied from item level, but also from bundle and test level, which means future studies should give more attention and concern to DBF (Differential Bundle Functioning) and DTF (Differential Test Functioning).

REFERENCES

[1] Bond, T. G., & Fox , C. M.. Applying the Rasch model: Fundamental measurement in the human sciences. New York: Routledge, 2015.

[2] Dong, S., & Ma, S.. The Comparative Research of the Three Usual Detect Procedure of DIF. Psychological Exploration In China, (1), 43-48, 2001.

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

[4] Drasgow, F.. Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19–29, 1987.

[5] Geranpayeh, A., & Kunnan, A. J.. Differential item functioning in terms of age in the certificate in advanced English examination. Language Assessment Quarterly, 4 (2), 190-222, 2007.

[6] Holland, P. W., & Thayer, D. T.. Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum, 1988.

[7] Ibrahim, A.. Differential item functioning: The state of the art. Jigawa Journal of Multidisciplinary Studies (JJMS), 1(1), 37-50, 2018.

[8] Jensen, A.R. 1980: Bias in mental testing. New York: Free Press.

[9] Kim, M.. Detecting DIF across the different language groups in a speaking test. Language Testing, 18, 89–114, 2001.

[10] Kunnan, A. J..Test Fairness. In M. Milanovic & C. Weir (Eds.), European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference(pp. 27-48). Cambridge: Cambridge University Press, 2004.

[11] Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S.. Differential Performance on National Exams: Evaluating Item and Bundle Functioning Methods using English, Mathematics, and Science Assessments. Sage Open, 6(2): 1-14, 2016.

[12] Lei, X.. Gender Difference in Achievements of English Subtest of Shanghai College Entrance Examination and Its Causes. Examination and Evaluation In China, (6), 43-46,2007.

[13] Li, H. & Suen, H. k.. Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. Language Testing, 30(2), 273–298, 2012.

[14] Li, Q. & Kong, W.. DIF Study of Reading Module in TEM-4. The Journal of Foreign Languages in China, 6(1), 53-60, 2009.

[15] Lord, F. M.. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

[16] Mellenbergh, G. J.. Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118, 1982.

[17] Pae, T.. DIF for examinees with different academic backgrounds. Language Testing; 21, 53-73, 2004.

[18] Pae, T., & Park, G.. Examining the relationship between differential item functioning and differential test functioning. Language Testing, 23 (4) 475–496, 2006.

[19] Raju, N. S.. The area between two item characteristic curves. Psychometrika, 53, 495-502, 1988.

[20] Ryan, K., & Bachman, L.. DIF on two tests of EFL proficiency. Language Testing, 9, 12–29,1992.

[21] Salehi, M., & Tayebi, A.. Differential Item Functioning: Implications for Test Validation. Journal of Language Teaching and Research, 3(1): 84-92, 2012.

[22] Shealy, R., & Stout, W.. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test/DTF as well as item bias/DIF. Psychometrika, 58, 159-194, 1993.

[23] Swaminathan, H., & Rogers, H. J.. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361–370, 1900.

[24] Uiterwijk, H., & Vallen, T.. Linguistic sources of item bias for second- generation immigrants in Dutch tests. Language Testing, 22, 211–234, 2005.

[25] Zumbo, B. D.. Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. Language Testing, 20, 136–47, 2003.

[26] Zwick, R., Donoghue, J. R., & Grima, A.. Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30, 233- 251, 1993.