

Using Global and Local DIF Analyses to Assess DIF across Language Groups

Xuan Tan

Mark J. Gierl

Centre for Research in Applied Measurement and Evaluation

University of Alberta

6-110 Education North

Edmonton, AB, Canada T6G2G5

Abstract

Most statistical procedures for identifying differential item functioning (DIF) share one important practice in computing the DIF measure: They indicate the magnitude of DIF by calculating the *average difference across all ability levels*. However, the averaging effect weakens the power of these procedures in detecting some unique and noteworthy types of DIF such as crossing DIF or DIF occurring only within a small range of ability levels. In this paper, three other test statistics, Fisher's χ^2 , Cochran's Z, and Goodman's U (Marascuilo & Slaughter, 1981), are used with the aid of graphical DIF procedures for local DIF analysis to study DIF that occurs in specific ranges of an ability level. The usefulness of local DIF analysis is demonstrated by its enhancement in interpreting DIF and its potential application in criterion-referenced tests.

Using Global and Local DIF Analyses to Assess DIF across Language Groups

In the process of establishing the validity for specific test score inferences, one important step is to assess the fairness of the test through the analysis of differential item functioning (DIF). The *Standards for Educational and Psychological Testing (Standards)* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999) stated, “The term bias in tests and testing refers to construct-irrelevant components that result in systematically higher or lower scores for identifiable groups of examinees.” DIF is used to detect group differences by assessing the probability that individuals of equal ability would answer an item correctly, which is exactly what the *Standards* identified as an issue that should be addressed in test development.

Ample research has been conducted to assess DIF across gender, ethnicity, and language groups (e.g., Hamilton, 1999; Zhang, 2002; Gierl, Bisanz, Bisanz, & Boughton, 2003). A number of statistical procedures used commonly today include the SIBTEST procedure (Shealy & Stout, 1993a), the Mantel-Haenszel statistic (Holland & Thayer, 1988), comparison of item response functions (Thissen, Steinberg, & Wainer, 1993), and logistic regression (Swaminathan & Rogers, 1990). Research has also been conducted to employ different types of analyses to try to reduce bias and detection error through iterative DIF procedures (see Zenisky, Hambleton, & Robin, 2003). These global DIF procedures (DIF analysis across the entire ability scale) all share one important practice in computing the DIF measure: They indicate the magnitude of DIF by calculating *average difference* across all ability levels. In cases where DIF is present only in a specific range of ability or where the direction of DIF changes across ability levels, the averaging effect weakens the power of these procedures in detecting these unique and noteworthy types of

DIF. Li and Stout (1996) developed a method called Crossing SIBTEST that can be used to assess crossing DIF (the direction of DIF changes across ability levels). This procedure was shown to be more powerful than SIBTEST and Mantel-Haenszel in detecting crossing DIF (Li & Stout, 1996). However, the DIF statistic in this procedure is calculated with only one crossing point near the middle of the score scale. When there are multiple crossing points or the crossing point is farther from the middle of the ability scale, Crossing SIBTEST is no better than the other DIF procedures in detecting DIF. Moreover, the averaging effect still exists, which weakens the statistical power for detecting DIF when it occurs only within a small region on the ability scale.

To overcome the limitations associated with different statistical procedures, Bolt and Gierl (2004) suggested using three other test statistics—Fisher's χ^2 , Cochran's Z, and Goodman's U test—for local DIF analysis (DIF analysis focusing on specific ability levels of interest). These three test statistics were shown to be more powerful in detecting DIF with changing direction and magnitude. Further, these test statistics, when combined with graphical DIF analysis, could shed more light on investigating the causes of DIF. After DIF items with changing direction and magnitude are identified, graphs of item response functions (IRFs) can be used to identify ability ranges of interests (within which DIF is significantly large or between which the direction of DIF changes). Protocol analysis can then be followed to identify causes of DIF by interviewing only students of the interested ability ranges (Ericsson & Simon, 1984; Leighton, 2004). By excluding students of ability ranges where there is no DIF, the interpretability of the DIF will likely improve for two reasons. First, by excluding students from the comparing groups who did not perform differently on an item, the sample of students interviewed will only include students who can provide useful information on differential performance. Substantive analyses with interview data will therefore be more focused. Second,

the sample size required for interview can be reduced because only students in a specific ability range are needed. Reduced sample size will make both the interview process and the analysis process less cumbersome and, likely, more informative.

In Bolt and Gierl's (2004) paper, the precision of the three test statistics with regression correction implemented (a procedure for correcting the bias in using the total test score as a surrogate for ability) was studied. But then simulations were focused on global DIF analysis. Local DIF analysis was proposed as a promising new direction. However, much research is still needed to carry out the application of this newly proposed method to evaluate its feasibility and usefulness. Thus, the present study is designed: (1) to evaluate performance difference between the English- and French-speaking students using both global and local DIF analyses to see how the DIF procedures compare and contrast and (2) to illuminate the potential usefulness of local DIF results for studying translation DIF.

Method

Data

Data from the 1997 administration of the Alberta achievement test for Grade 9 Mathematics were used for this study. Achievement tests are given to Alberta students in all core courses in Grade 3, 6 and 9 to ensure high standard of education and accountability from both teachers and students. The Grade 9 mathematics test covers the content areas of number concepts and operations, patterns and relations, shape and space, and statistics and probability across two thinking levels—knowledge and skills. The total number of items on the test is 55. The target population of the test is all Grade 9 students in Alberta. Data from the 1997 administration include all Grade 9 students of that year who took the Mathematics achievement test.

Language DIF has always been a major concern on bilingual standardized tests. With two languages of instruction existing in Alberta, English and French, it is essential to not only develop the achievement tests in two languages, but also ensure the equivalence of the two tests so neither of the language groups is disadvantaged due to item bias (Sireci, 1997). Language was selected as the grouping variable in this study because of this concern.

The data included student responses to the test across two language groups, English and French. The French group included French-speaking students and French immersion students. Students in the French group took the same tests as the English-speaking students except that the test items were translated into French. Six items were not included in the French version. Therefore, these items were deleted from the English version producing a final dataset with 49 items for this study. Forty of the items were multiple-choice items and nine items were numerical-response items. All items were dichotomously scored. The final sample sizes were 3000 and 2115, respectively, for the English and the French group. English is the reference group while French is the focal group.

The Regression Correction Procedure to Remove Bias in DIF Analyses

DIF is measured by comparing IRFs across groups. An IRF is a logistic function that represents the probability of a correct response to an item as a function of the latent ability (θ). The DIF function can be expressed as $B(\theta) = P_F(\theta) - P_R(\theta)$, where $P_F(\theta)$ is the probability of correct response given θ for the focal group and $P_R(\theta)$ is the probability of correct response given θ for the reference group. $B(\theta)$ is summarized across ability levels, and then global DIF measures are obtained by averaging $B(\theta)$ over θ . A common strategy used to calculate the DIF index is to use total test scores as surrogates for θ . This strategy introduces problems and concerns when the groups to be compared have different latent ability distributions. Due to the

existence of measurement error and differential ability distributions, the same observed test score will not represent examinees of comparable ability across groups. If one group has higher mean ability than the other group, the same observed score will constantly produce a higher estimated true score for that group. When this happens, items showing no DIF will be incorrectly identified as having item bias.

In order to control the bias brought by using total scores as surrogates for θ , Shealy and Stout (1993b) proposed a regression correction procedure to reduce the Type I error rate (flagging items as DIF items when actually they are not). Examinees are first matched on valid subtest scores. The valid subtest contains items that are believed to be free of bias and are measuring the latent trait that is intended to be measured. A valid subtest score is sometimes calculated by subtracting the score of the item being studied from the total score. The proportion correct scores of the reference and focal group having the same valid subtest score k , p_{Rk} and p_{Fk} , are adjusted using regression correction to two new values, p_{Rk}^* and p_{Fk}^* . The adjusted values are closer to what the estimates should be for examinees with the same ability. Then, the difference between the two adjusted proportion correct scores is used to evaluate DIF.

In order to calculate the regression corrected proportion correct score estimates, the valid subtest score means, variances, and reliabilities for the reference and focal group need to be obtained first. Estimated true scores $\hat{V}_R(k)$ and $\hat{V}_F(k)$ can then be computed using these values as a function of the observed test scores in each group (Crocker & Algina, 1985, p. 147). The mean of the two true scores $\hat{V}_R(k)$ and $\hat{V}_F(k)$ is denoted as $\hat{V}(k)$. Then, the estimate of the slope (\hat{M}_{gk}) of the IRF in the region of score k for group g is calculated using the formula:

$$\hat{M}_{gk} = \frac{P_{g,k+1} - P_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)}.$$

The adjusted proportion correct score for each group is calculated as

$$p_{gk}^* = p_{gk} + \hat{M}_{gk}(\hat{V}(k) - \hat{V}_g(k)).$$

The regression correction procedure is embedded in the SIBTEST (Shealy & Stout, 1993b) and Crossing SIBTEST (Li & Stout, 1996) procedures.

SIBTEST

SIBTEST (Stout & Roussos, 1996) is a nonparametric method for assessing DIF/DBF (differential bundle functioning) based on Shealy and Stout's (1993a) multidimensional model for DIF. The DIF index, $\hat{\beta}_{UNI}$, is the weighted expected score difference between the reference and focal group of the same ability. $\hat{\beta}_{UNI}$ is computed as

$$\hat{\beta}_{UNI} = \sum_{k=0}^N \hat{f}_k (p_{Rk}^* - p_{Fk}^*),$$

where N is the maximum subtest score and \hat{f}_k is the weight applied to the difference in proportion correct scores for the reference and focal group with valid subtest score k .

The Three Test Statistics and TESTGRAF

Fisher's χ^2 test. Marascuilo and Slaughter (1981) proposed Fisher's χ^2 to test if the difference between the IRFs of two groups at one specific ability level was statistically significant. The Fisher's test evaluates the statistical dependence between group membership (reference/focal) and item response (correct/incorrect) conditioned on valid subtest score level k (Bolt and Gierl, 2004). The formula for the test statistic with regression correction applied is:

$$\chi_k^2 = \frac{(N_{Rk} + N_{Fk} - 1)(n_{Rk0}^* n_{Fk1}^* - n_{Rk1}^* n_{Fk0}^*)^2}{(n_{Rk0}^* + n_{Rk1}^*)(n_{Rk0}^* + n_{Fk0}^*)(n_{Rk1}^* + n_{Fk1}^*)(n_{Fk0}^* + n_{Fk1}^*)},$$

where N_{Rk} and N_{Fk} denote the total number of examinees having valid subtest score k in the reference and focal group respectively, n_{Rk1}^* and n_{Fk1}^* denotes the regression-corrected number of

examinees (frequencies) in the reference and focal groups who obtained valid subtest score k and answered the item correct, and n_{Rk0}^* and n_{Fk0}^* denotes the regression-corrected number of examinees in each group who answered the item incorrect. The regression-corrected frequencies can be calculated using the total number of examinees multiplied by the adjusted conditional proportion correct or incorrect scores obtained from the SIBTEST extended output. The test statistics can only be calculated when $N_{Rk} + N_{Fk} \geq 20$ and each of n_{Rk1}^* , n_{Fk1}^* , n_{Rk0}^* , and n_{Fk0}^* is at least 3 for χ_k^2 to approximate the chi-square distribution with 1 degree of freedom (Kanji, 1993).

The χ_k^2 statistic at each valid subtest score level can be summed up to produce an omnibus test of DIF across a range of ability levels:

$$\chi_T^2 = \sum_{k \in \kappa} \chi_k^2,$$

where κ includes all valid subtest scores that are of interest. The resulting statistic approximates the χ^2 distribution with degrees of freedom equal to the number of valid subtest scores in κ .

Fisher's test is most powerful in detecting significant DIF with changing directions. It can identify DIF no matter how many crossing points there are on the IRFs. However, the test is weaker when DIF is unidirectional across ability levels (Marascuilo & Slaughter, 1981).

Cochran's Z test. Cochran's Z test is more powerful than Fisher's χ^2 test when DIF is unidirectional. It evaluates the null hypothesis $B(\theta) = c$, where c is any non-zero constant. The difference between IRFs at any valid subtest score k , \hat{B}_k , is denoted

$$\hat{B}_k = p_{Rk}^* - p_{Fk}^*,$$

where p_{Rk}^* and p_{Fk}^* are the regression corrected proportion correct scores. The average difference between IRFs, \hat{B}_0 , can be then computed as

$$\hat{B}_0 = \frac{\sum_k^N W_k \hat{B}_k}{\sum_k^N W_k},$$

where

$$W_k = \frac{N_{Rk} N_{Fk}}{N_{Rk} + N_{Fk}},$$

is the weight associated with each valid subtest score. The final test statistics is given by

$$Z = \frac{\hat{B}_0}{SE_{B_0}},$$

where

$$SE_{B_0}^2 = \frac{1}{(\sum_k^N W_k)^2} \sum_k^N W_k \frac{N_{Rk}^2 p_{Rk}^* (1 - p_{Rk}^*) + N_{Fk}^2 p_{Fk}^* (1 - p_{Fk}^*)}{(N_{Rk} + N_{Fk})^2}.$$

Under the null hypothesis, Z has a standard normal distribution. Cochran's Z is similar to the $\hat{\beta}_{UNI}$ of SIBTEST. However, the weighting functions are different for the statistics.

SIBTEST uses either the focal group valid subtest score density or a pooled reference and focal group density as weights (Shealy & Stout, 1993a), while Cochran's Z uses W_k . The different weighting functions determine that Cochran's Z has more power in detecting DIF while the $\hat{\beta}_{UNI}$ of SIBTEST yields a more interpretable DIF measure because it can be interpreted as an expected conditional difference between the reference and focal groups on the item in reference to a test score distribution in a known population.

Goodman's U test. Goodman's U test evaluates whether the amount of DIF in an item varies across ability levels. The formula for computing the final test statistics is:

$$U = \sum_k^N W_k (\hat{B}_k - \hat{B}_0)^2,$$

where

$\hat{B}_0 = \frac{\sum_k^N W_k \hat{B}_k}{\sum_k^N W_k}$ is the average difference between IRFs across all valid subtest scores with \hat{B}_k

defined the same way as for the Cochran's Z test,

$(\hat{B}_k - \hat{B}_0)$ is the displacement of \hat{B}_k from \hat{B}_0 ,

$W_k = \frac{1}{SE_{\hat{B}_k}^2}$ is the weight to be applied to the displacement quantity, and

$SE_{\hat{B}_k}^2 = \frac{P_{Rk}^*(1 - P_{Rk}^*)}{N_{Rk}} + \frac{P_{Fk}^*(1 - P_{Fk}^*)}{N_{Fk}}$ is the error variance.

The same decision rule as the Fisher's χ^2 test is used for inclusion of a valid subtest score category in the computation of Goodman's U test (i.e., $N_{Rk} + N_{Fk} \geq 20$, and each of n_{Rk1}^* , n_{Fk1}^* , n_{Rk0}^* , and n_{Fk0}^* is at least 3). Under the null hypothesis of consistent DIF across ability levels, U approximates the χ^2 distribution with degrees of freedom equal the number of valid subtest score categories used in the computation of U. An item with a significant U test would be a candidate item for local DIF analysis.

TESTGRAF. TESTGRAF (Ramsay, 2000) is a computer program that can produce the graphical representation of the reference and focal group IRFs. TESTGRAF uses kernel smoothing to obtain nonparametrically estimated IRFs that can be used to compare groups. Although the statistical significance of the difference between IRFs between two groups cannot be evaluated, the graphs produced by TESTGRAF provide a starting point for further analysis based on specific features of the graphs (i.e., varying amount of DIF implying local DIF analysis, or consistent amounts of DIF implying global DIF analysis).

Data Analysis

Bolt and Gierl (2004) applied the regression correction procedure, which is embedded in the SIBTEST procedure, to the three DIF tests—Fisher’s χ^2 , Cochran’s Z, and Goodman’s U test. Using simulated data, they found that the Type I error rate was improved using regression correction and, when combined with graphical DIF analysis, that these three statistics were informative for describing the characteristics of the global DIF. The present study applied their method to the Grade 9 Mathematics achievement test data using English and French examinees to study global and local translation DIF.

Global DIF analysis was conducted first. SIBTEST (Stout & Roussos, 1996), as a conservative DIF test with regression correction embedded, was used in the initial step to get an overall DIF measure. The extended output was produced to aid the calculation of the three test statistics. Global DIF indices were calculated for the Fisher’s χ^2 , Cochran’s Z, and Goodman’s U tests to see how these results compare and contrast with the SIBTEST results. TESTGRAF was then used to see if local DIF was evident in the IRF curves. Items with unique patterns of DIF and significant U tests were identified, and the IRF curves were produced to select the specific ability levels of interest for local DIF analyses.

Results*Global DIF Results*

The $\hat{\beta}_{UNI}$ statistics for the math items were calculated using SIBTEST, and then the items were classified using Roussos and Stout’s (1996) guidelines: (1) negligible or A-level DIF: Null hypothesis is rejected and $|\hat{\beta}_{UNI}| < 0.059$, (2) moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\hat{\beta}_{UNI}| < 0.088$, and (3) large or C-level DIF: $|\hat{\beta}_{UNI}| \geq 0.088$. Of the 49 items,

40 items displayed A-level DIF (i.e., no DIF), 4 items displayed B-level DIF, and 5 items displayed C-level DIF.

The item response function (IRF) curves of all items were produced using TESTGRAF. As one might expect from the SIBTEST result, only the IRFs of the 9 DIF items should have notable discrepancies. However, this was not the case. Most of the IRFs showed some level of discrepancy between the two groups indicating the possible existence of DIF. Also, the IRFs of some items were inconsistent across ability levels. Although these graphs are very informative as to specific characteristics of DIFs, they do not provide significance test. Therefore, the three test statistics were used to see if the items shown in the graphs displayed significant group differences.

Results from the significance tests of the Fisher's χ^2 , Cochran's Z, and Goodman's U for all items across the entire ability scale are included in Table 1. The items are categorized according to the SIBTEST results. As we can see from the table, for the B and C-level DIF items flagged by SIBTEST, the χ^2 and Z tests were also significant. The U test was not significant for the nine DIF items indicating the amount of DIF across ability levels is consistent. The IRFs for item 25 (B-level DIF) and item 15 (C-level DIF) are also included in Figure 1. Consistent discrepancies between the IRFs across the ability scale are apparent confirming the significant χ^2 and Z tests.

However, for the 40 A-level items, the significance tests of the three test statistics contradicted some of the SIBTEST results. The three test statistics were not significant for 11 of the 40 A-level items. However, for the other 29 items, at least one of the test statistics was significant. These items are separated by the dash lines in Table 1. There are items identified by all three tests, items identified by two of the three tests, and items identified by only one of the

three tests. Inspection of the IRF graphs of these 29 items showed different types of discrepancies between the two groups. For illustration purpose, one item was taken from each of the categories separated by dash lines. Items 12, 16, 23, and 7 are presented in Figure 1.

Item 12 represents the two items with significant results in all three test statistics. In its IRF curves, there are several notable areas with wide gaps between the IRFs of the two groups. The direction of DIF is fairly consistent favoring the focal group. The fact that the gap between the two IRF curves varies dramatically in width across ability levels likely produced the significant U test.

Item 16 represents the three items with significant χ^2 and Z test results. Both tests evaluate the magnitude of DIF, and are more powerful than the $\hat{\beta}_{UNI}$ test in SIBTEST. In the IRFs graph for item 16, there is a fair amount of discrepancy between the two IRF curves consistently favoring the focal group thereby suggesting this is indeed a DIF item.

Item 23 represents an item with significant Z test results. The Cochran's Z test is most powerful when DIF is unidirectional. The IRF curves for item 23 clearly demonstrate a consistent but narrow gap between the two groups. The Fisher's χ^2 test is not powerful enough to identify this weaker DIF item. Comparing the IRFs graphs for item 16 and 23, we can see item 16 has more DIF than item 23.

Item 7 represents the two items with a significant U test meaning the amount of DIF varies across ability levels. The IRF curves of item 7 have several crossing points, and the gap between the two curves fluctuates noticeably at several locations on the score scale. This item response pattern accounts for the significant U test.

Local DIF Results

Items 12 and 7 were further investigated because they represented the items having a significant U test. A significant Goodman's U test indicates that the amount of DIF varies across ability levels. As seen in Figure 1, both of the items have a gap of varying width between the IRFs. However, for item 12 which has significant results for all the three tests, the discrepancy between the IRFs is fairly large around some ability levels whereas, for item 7, which has only one significant result for the U test, the discrepancy between the IRFs is relatively small. Thus, item 12 is a good candidate for local DIF analysis to see if DIF exists in a specific range of ability.

For item 12, notable DIF occurred around the ability range of -1.6 to 0.2 and the ability range of 0.7 to 1.7 . The Fisher's χ^2 and Goodman's U test were calculated around these two specific ranges of ability levels. Results of the local DIF analysis are presented in Table 2. The χ^2 test is significant for both ability ranges, while the U test is significant only for the ability range of -1.6 to 0.2 . This outcome is also evident from the graph: The amount of DIF for the ability range of -1.6 to 0.2 fluctuates several times along the ability range whereas the discrepancy between the IRFs for the ability range of 0.7 to 1.7 is fairly consistent, as there is not much fluctuation between the IRFs. Thus, there is local DIF in both ability ranges, but DIF within the ability range of -1.6 to 0.2 maybe attributable to only several ability levels within that range.

Discussion and Educational Implications

The difference between the results obtained from SIBTEST alone and from TESTGRAF combined with the three test statistics draws attention to the use of alternative methods for detecting DIF. The three test statistics—Fisher's χ^2 , Cochran's Z, and Goodman's U—are

shown to be more powerful than the $\hat{\beta}_{UNI}$ statistics of SIBTEST in detecting DIF *in some situations*. Items 12, 16, 23, and 7 represent items that had significant results with some or all of the three test statistics but were not identified as showing DIF by SIBTEST. However, inspection of the IRFs graphs of these items revealed that there was notable discrepancy in performance between the reference and focal groups. Due to limited resources available to this project, we only studied the statistical method. No evaluation of item features that may produce these results was carried. Therefore, further studies that examine the characteristics of the items along with the DIF results may prove useful in providing further information on whether DIF actually exists in those items identified by the three test statistics but not by SIBTEST.

In sum, the most important finding in this study is that the three statistics can be combined with the IRF graphs to study local DIF in a small range of the ability score scale which, in turn, may help us identify the causes of DIF. With significant DIF results found around the specific ability levels for item 12, the item can be further studied by examining the responses of students at those specific ability levels. Protocol analysis can also be carried out by interviewing students at those specific ability levels. The merit of local DIF procedures lie in the possibility of identifying specific features of the item that might cause students at *specific ability levels* to perform differently from the others. This information could contribute, in part, to the interpretation problems that often hamper DIF research (e.g., Englehard, Hansche, & Rutlege, 1990; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999).

Another important and promising application of local DIF analysis is its use in studying DIF on criterion-referenced tests. For a criterion-referenced test, setting valid cut-scores is crucial. Misclassification rate must be minimized. Items showing DIF at the ability levels around the cut-off points would be devastating because the possibility of misclassifying people

increases. In this case, local DIF analysis that concentrates on the ability levels of interests (around the cut-off points) could bring new insight into the validation process and contribute to the establishment of the cut-scores. Interview studies can also focus on these specific ability levels based on local DIF results.

Despite the apparent appeal of local DIF analysis in different situations, we should be aware of the procedure's weaknesses. The three test statistics—Fisher's χ^2 , Cochran's Z, and Goodman's U—have improved control of Type I error when regression correction is used (Bolt & Gierl, 2004). However, they are still susceptible to problems associated with sample size fluctuations. When sample size increases to 1000, there is a small but notable inflation of Type I error rates, especially when the ability distributions of the groups are different. Effect size measures for the three test statistics need to be developed to ensure that statistically significant results are meaningful and not caused by large sample size. The Cochran's Z test may also be overly liberal as it tends to overflag items, given 2/3 of the items on the Mathematics 9 test were flagged as showing DIF by the Z test. This outcome makes the result of the Z test less credible. Therefore, we recommend that the result from the Z test should always be examined together with the other two tests when making a decision about an item. Finally, the use of TESTGRAF to locate the ability ranges where local DIF occurs still needs to be evaluated. The graphs TESTGRAF produces are influenced by the actual student responses. Thus, the graphs are just a nonparametric representation of the IRF curves. Decision about the starting and ending points of the ability ranges is typically judgmental. All these limitations associated with the local DIF procedure should be recognized when interpreting local DIF results. Local DIF analysis is a promising direction for DIF research. But its potential applications should be fully studied to ensure we make the best use of this new DIF approach.

References

- Bolt, D. M., & Gierl, M. J. (2004, April). *Application of a regression correction to three nonparametric test of DIF: Implications for global and local DIF detection*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Crocker, L. & Algina, J. (1985). *Introduction to classical and modern test theory*. Orland, FL: Harcourt Brace Jovanovich.
- Englehard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 3, 347-360.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal Reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Gierl, M. J, Bisanz, J., Bisanz, G. L, & Boughton, K. A. (2003) Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, 40, 281-306.
- Gierl, M. J, Bisanz, J., Bisanz, G. L, Boughton, K. A, & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.
- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and substantive reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research*, 45, 353-376.

- Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*, 211-235.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ.
- Kanji, G. K. (1993). *100 statistical tests*. Thousand Oaks, CA: Sage.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6-15.
- Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.
- Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18*, 229-248.
- Ramsay, J. O. (2000). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data*. Department of Psychology. McGill University.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.
- Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.
- Shealy, R. T., & Stout, W. F. (1993a). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 197-239). Hillsdale, NJ.

- Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true-bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *54*, 159-194.
- Stout, W., & Roussos, L. (1996). *SIBTEST manual*. Statistical Laboratory for Educational and Psychological Measurement. University of Illinois at Urbana-Champaign.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ.
- Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*, 51-64.
- Zhang, Y. (2002, April). *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.

Table 1

Significance Test of the SIBTEST and the Three Test Statistics for the Grade 9 Math Test

SIBTEST Result	Item Number	Fisher's χ^2	Cochran's Z	Goodman's U
	12	*	*	*
	13	*	*	*
	16	*	*	
	41	*	*	
	46	*	*	
	1		*	
	3		*	
	8		*	
	9		*	
	11		*	
	14		*	
	17		*	
	18		*	
	19		*	
	20		*	
	21		*	
	23		*	
	24		*	
	28		*	
A-level DIF	30		*	
	31		*	
	32		*	
	34		*	
	35		*	
	38		*	
	42		*	
	45		*	
	5			*
	7			*
	4			
	6			
	10			
	22			
	29			
	33			
	39			
	43			
	44			
	47			
	49			
B-level DIF	2	*	*	
	25	*	*	
	37	*	*	
	40	*	*	
C-level DIF	15	*	*	
	26	*	*	
	27	*	*	
	36	*	*	
	48	*	*	

* $p \leq 0.05$

Table 2

Local DIF Results—Significance Test of the Fisher's χ^2 and Goodman's U for Item 12

Range of Ability	Fisher's χ^2 (df)	Goodman's U (df)
-1.6 to 0.2	24.84988* (13)	26.03215* (13)
0.7 to 1.7	17.83195* (9)	12.07261 (9)

* $p \leq 0.05$

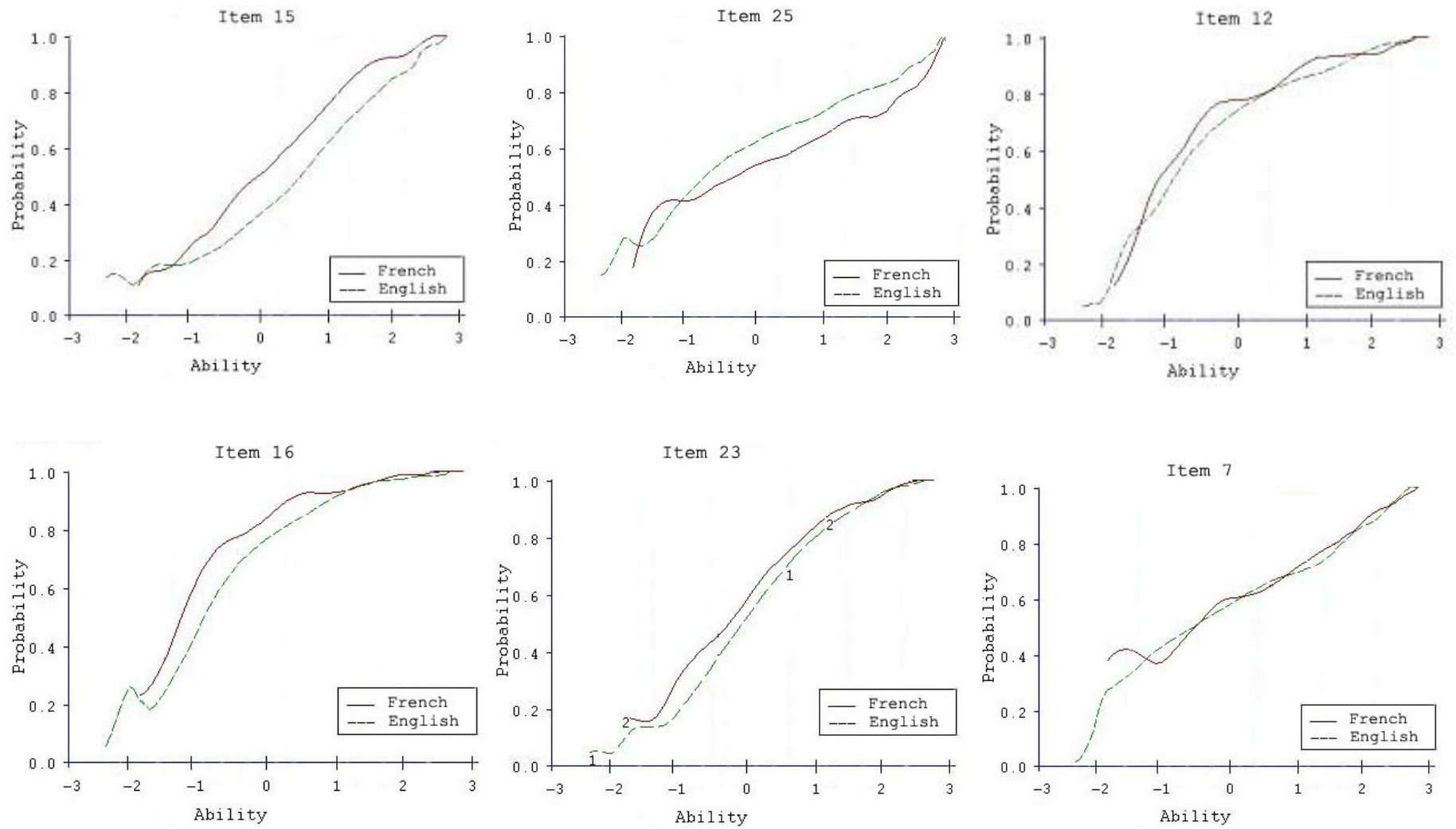


Figure 1. Item Response Function Curves of the Four Representative Items