# Differential Item Functioning Analyses with STDIF: User's Guide

April L. Zenisky, Frédéric Robin, and Ronald K. Hambleton

*[Version 6/15/2009]*

## Part I: Introduction to the Mechanics of SDIF and UDIF

STDIF is a DOS-based program written by Frédéric Robin (2001) to compute DIF indices of conditional p-value differences between two groups of interest: the reference group and the focal group. This is a large-sample procedure requiring a minimum sample size of 10 people *in each group* (the reference and focal group) at each score point, and was designed to be used with state level data, not pilot samples where sample sizes are typically smaller.

This program actually computes two different indices of DIF: *SDIF* and *UDIF*. It can handle datasets including up to 500,000 examinees and 150 items, and total test scores up to 200 points.

**The SDIF index:** The SDIF (signed DIF) index expresses the signed weighted average difference between reference and focal group conditional p-values, and is a statistic calculated for each item on the test to provide a single number for flagging DIF items (Dorans & Kulick, 1986). It is computed as:

$$SDIF = \sum_{s=0}^{K} w_s (p_s^R - p_s^F)$$

where K is the maximum number of score points that a student can achieve on the test; $p_S^f$ is the proportion correct score for the focal group who received a test score of *s* (i.e., rescaled p-value conditioned on *s*); similarly, $p_s^r$ is the conditional p-value for members of the reference group who received a test score of *s*; and $w_S$ is the standardization weight at each score level *s*. In that this index allows for reference and focal group p-values to cancel each other out, the statistic only provides insight into levels of uniform DIF.

A note on **standardization weights**, $w_s$, in the SDIF (and UDIF) statistics: There will be occasions when the researcher wants differences between reference and focal groups at each score point to count equally in the calculation of DIF. More often, the choice is to have the weights reflect the proportion of total candidates (reference plus focal) at each score point. Finally, at other times, the main interest is in the focal group only (often this is the case when doing Black/White or Hispanic/White DIF studies). In this situation, the researcher wants to weight any reference-focal group item performance difference at a score point by the proportion of the focal group who are at that score point.

All of these options are available in STDIF. As indicated in Part II below (Generic example), choosing the "Weighting of cases" code of: "0" will produce an SDIF statistic in which there is no weighting at all, "1" will result in a weight at each score point corresponding to the proportion of both reference and focal group members, and finally, "2" will result in weighting

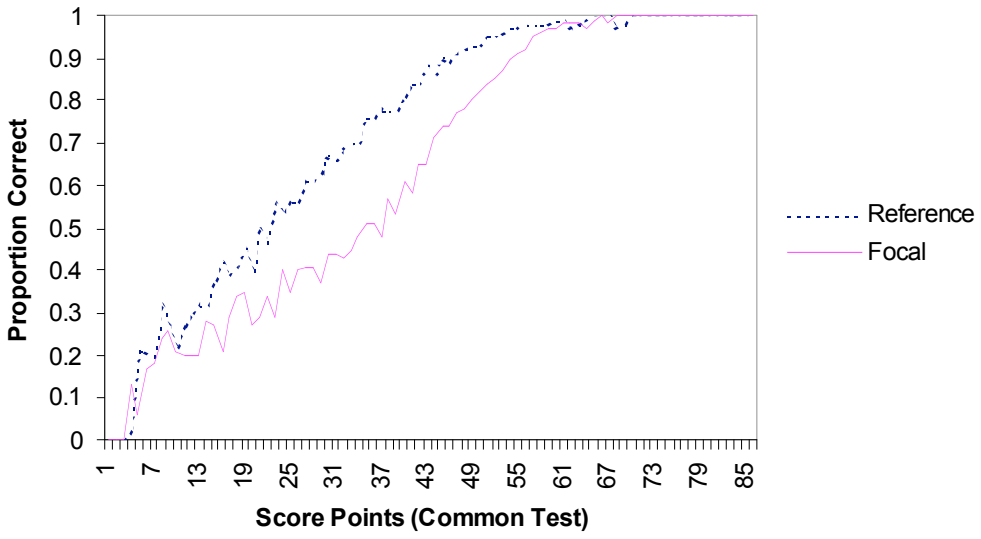
the conditional difference at a score point by the proportion of focal candidates who are at that score point.

**The UDIF index**: The UDIF (unsigned DIF) index is very similar to the SDIF index except that it provides a means for gauging the magnitude of differences between item p-values for members of the reference group and the focal group where both uniform and non-uniform DIF is present. It reflects the absolute area between reference and focal conditional expected responses, and is computed as:

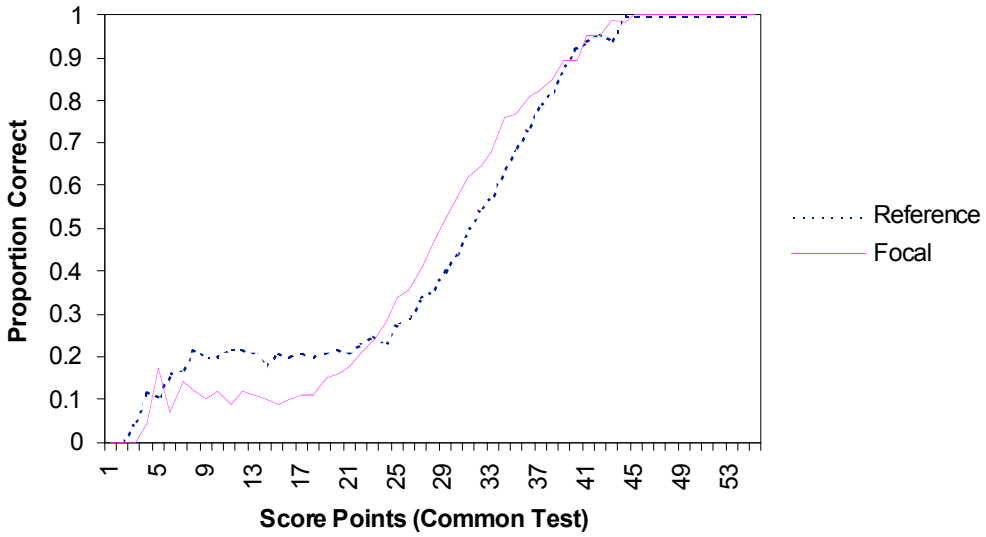
$$UDIF = \delta \sum_{s=0}^{K} w_s \left| p_s^R - p_s^F \right|$$

where $\delta$ is set to +1 if the item favors the reference group and to –1 otherwise. The only value of δ is to provide information about the direction of the DIF. (UDIF will always have a greater value than SDIF except in one instance: The two statistics will be equal when the p-value differences between groups at each score point are consistently in the same direction or zero. In our own research we have tended to use the UDIF statistic as the more important of the two for flagging DIF. When the statistics are very different in value, non-uniform DIF is the cause.)

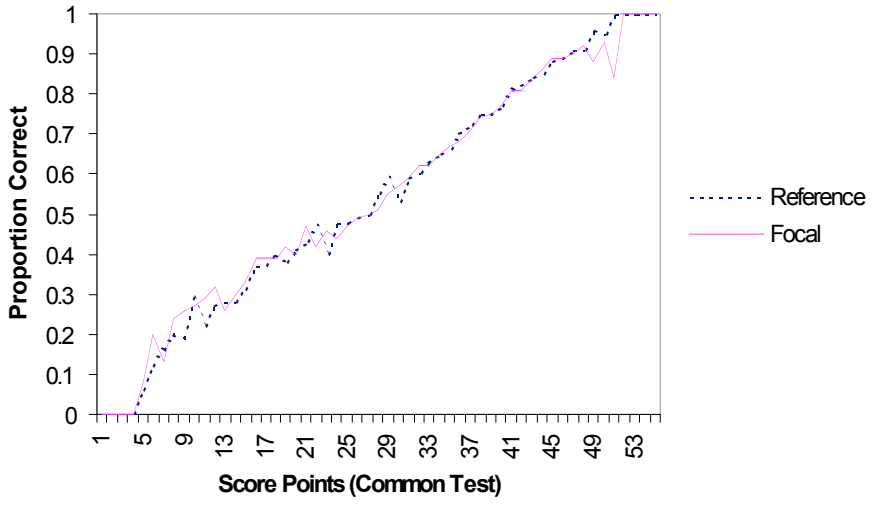
**Uniform and non-uniform DIF**: Uniform DIF refers to situations where the differences between reference and focal group p-values are relatively constant across different points in the examinee ability distribution. The graph below depicts uniform DIF (SDIF=0.135, UDIF=0.136).



Non-uniform DIF reflects instances where the reference group outperforms the focal group in one part of the ability distribution, and in another part of the distribution the opposite is true (the relative proficiency of reference vs. focal group examinees seems to switch as ability increases). This next graph shows non-uniform DIF, and the DIF here is small, as the differences in performance between the two groups are *on average* not large, although clear differences are present at different points in the score scale (SDIF=0.019, UDIF=0.040).




For comparison purposes, the graph below represents an item where DIF is not present. (SDIF=0.001, UDIF=0.017)



## Part II: Carrying out the Analyses

To do these analyses, you need Robin's (2001) STDIF program, a command file, and a data file (text).

- The program is available as shareware. As it is a DOS program, to run it is a matter of typing 'stdif *filename*.cmd' at the DOS prompt.

- The command file can be created in DOS or in any text editing program (Notepad works well for this purpose). It is only 10 lines long, but these are 10 very important lines.

*Generic example:*

| | |
|---|---|
| Title | *Name of analysis (BE DESCRIPTIVE) |
| Name of data file | **filename*.dat |
| Number of examinees | *Just a number (total N of examinees in data set) |
| Number of items | *Just a number |
| Position of group identifier in data file | *The number of the column in which STDIF will find the group ID code |
| Reference group identifier | *e.g., gender analysis, M; race analysis, W |
| Focal group identifier | *e.g., gender analysis, F; race analysis, B |
| Position of first item in data file | *Column number where response data starts |
| (a FORTRAN format statement) | *In parentheses, explains columns of items |
| Minimum number of matched examinees | *Just a number: set at 10 (at least 10 examinees must be in both reference and focal group at each score level to make comparisons) |
| Rescale (1) polytomous items to 0-1, or not (0) | * e.g., 0. |
| Weighting of cases | *Coding with 2 in the command file corresponds to weighting conditional difference at a score point by the proportion of focal candidates who are at that score point; Coding with 1 corresponds to weighting conditional difference at a score point by the proportion of reference and focal candidates of the combined sample at that score point; Coding with 0 corresponds to no weighting at all. |

*Here is a working command file:*

| | |
|---|---|
| Sample Test: Math Grade 4, Gender DIF analysis | *title |
| M004_s1.dat | *data file |
| 76784 | *76784 examinees |
| 39 | *39 items |
| 12 | *Column 12 has gender info |
| M | *M for males (reference group) |
| F | *F for females (focal group) |
| 13 | *Item responses start in column 13 |
| (39I1) | *39 one-column-long integer variables |
| 10 | *Min. N of matched examinees |
| 0 | *No rescaling of polytomous items |
| 1 | *Differences at score points weighted by the proportion of the combined sample at the score points |

To create a good, working command file, you **must check your data** to be sure of what columns the various pieces of information are in and to know how many examinee data records to read in.

The command file should be named in a descriptive way, including NO MORE THAN 40 CHARACTERS. For example, *M004gen*.cmd as well as the more descriptive Math_2000_grade_4_Gender_analysis.cmd could be used.

To run a DOS program, the data file must be labeled as *filename*.DAT. For your own ease of analysis, you should probably continue to be descriptive in naming these files.

**NOTE:** You must modify the data files once you have made them into .DAT files. Using the DOS editor or a text editor program (probably Wordpad as the files might be too big for Notepad) you must insert three lines at the top of the data file.

The **first** line you will enter is the maximum score for each test item.
The **second** line you enter is a sequence of 0-1 "switches" for including or excluding items from the analysis. An example from grade 4 Math Gender analyses is below.
The **third** line you enter is one that corresponds to aggregation of items. For the purpose of many analyses this row should be a line of zeroes.

```
            1111111111111111111111111111111144444
            1111111111111111111111111111111111111
            0000000000000000000000000000000000000
5050000474  1101011011101110010100001011001101113 13
5050000502  0010000100011000010101101010010000000
5050000577WM0101100100100100000000000000000012000
9002038068WM1111110101111111110010111110011003324 3
9002038228WF0100000100100100000110000000000011010
9002038710HF0000111111000000000001000000010011201 3
9002038827WF1000101111001110111000111111101111222 13
Etc.
```

**From the first line**, you see that most items except the last 5 are dichotomously scored (the maximum score is 1); the last 5 are polytomously scored and the maximum is 4.

**From the second line**, the fact that there is a 1 in each column means that every item on the test in included in the DIF analysis. These are the **switches** that are important is terms of the DIF procedure we are using.

**This idea of "switches" is important**. In DIF analyses, we try and evaluate the statistical characteristics of items across different groups. Rather than focus on "overall" item statistics, DIF techniques are conditional. As Dorans and Holland (1993) pointed out, "In contrast to impact, which often can be explained by stable consistent differences in examinee ability distributions across groups, DIF refers to differences in item functioning after groups have been matched with respect to the ability or attribute that the item purportedly measures" (p. 37). In DIF analysis, test-takers from different groups are matched on the psychological attribute measured, and the probability of differential responses across matched test-takers is evaluated. Items are considered to be functioning differentially across groups if the probability of a particular response differs significantly across test-takers who are equivalent (i.e., matched) on proficiency. The DIF analyses conducted used total test score to identify females and males who were "equal" with respect to the proficiency measured by each test.

Some researchers have criticized DIF results because oftentimes people use total test score as the ultimate criterion. This is a problem when DIF is present because DIF items introduce a bias in the matching variable and this makes it impossible to properly match examinees using the total test scores of the reference and focal groups. A common solution (as we are implementing here) is to turn the DIF analysis into a two-stage procedure. In the first stage, total score is used as the matching variable. In the second stage, items showing DIF at the first stage are removed from the matching variable.

**From the third line**, zeroes is every data column means that each item is considered separately and not added in with any other item. It is possible to aggregate items by placing a "1" in the column of each item to be included in the aggregation. Only one combination of aggregations is permitted per run (in other words, you can select multiple items to aggregate, but all of those items are aggregated into one large bundle of items).

## Program's Output

Two new output files should be created once the command "STDIF *filename*.cmd" has been executed (entered directly within an opened DOS Command Prompt window or by executing a batch file including the same command). These will have the same *filename* but their extension will be .LOG and .SDO.

The purpose of the LOG file is to allow the user to check that the command file has been properly interpreted and that data have been properly read in. The LOG file also provides useful feedback concerning examinees without proper group identifier, the number of items removed, a table of score frequencies, etc. Finally, it provides two DIF item tables sorted by SDIF or UDIF and one summary DIF table.

The purpose of the SDO file is to provide the final results in a simple format that can easily be used to create the tables and graphs such as those displayed in this guide using Microsoft Excel, for example.

**Method/Example**

What we will be doing is actually running Robin's (2001) STDIF program on each data set **TWICE**. In the first run-through, we include every common item (thus, insert a sequence of 1's is on the second line of the data set).

From the output of that first analysis (a file that ends in .SDO), we look at the UDIF index (Column 4) and identify those items that appear to be showing DIF. In these DIF analyses, those items with DIF statistics that are positive favor the reference group, while those with statistics that are negative favor focal group examinees. But the direction of the DIF from the first stage of the analysis is unimportant. What is important is that items showing DIF, positive or negative, are eliminated from the criterion to obtain a less biased criterion for matching reference and focal group members. Please use a > (+/-) .075 criterion to start. Make a note of the items that have a UDIF value exceeding .075 or -.075.

**Be careful**: As you look at the UDIF indices of the sample items, you should know if the SDIF and UDIF indices for the polytomous items are or are not on a 0-1 metric as those values for the dichotomous items are. If the choice was made not to rescale, then the indices should be divided by the maximum number of score points for the item to obtain an indication of the DIF on a "per point basis." So, for example, if the maximum number of score points is 4 and UDIF=.20, On a per point basis, the amount of DIF is about .05 and this is not large enough to worry about. Even though a UDIF value of .20 seems high, that difference is on a four point item, and so the level is actually relatively small. When the DIF is viewed like the binary scored items on the 0-1 scale, the DIF is actually quite small (only a difference of .05 for each scoring point). Now if on the same 4 point item and if UDIF=.50, then on a per point basis, the difference is .125 and this difference is substantial and should be very much a concern.

*Example*: At first glance, an item with a UDIF value of 0.16 should be flagged. However, if the item is polytomous, (as is item 39 below in the example), divide that UDIF value by 4 (its IMXS value) to get a revised UDIF value of .04. Thus this item WOULD NOT be flagged.

Here's an example of the UDIF values.

| imxs | Item | SDIF | UDIF | |
|---|---|---|---|---|
| 1 | 1 | 0.03 | 0.03 | |
| 1 | 2 | 0.02 | 0.02 | |
| 1 | 3 | 0.07 | 0.07 | |
| 1 | 4 | -0.02 | -0.03 | |
| 1 | 5 | 0.00 | 0.02 | |
| 1 | 6 | -0.01 | -0.03 | |
| 1 | 7 | -0.03 | -0.03 | |
| 1 | 8 | -0.01 | -0.02 | |
| 1 | 9 | 0.01 | 0.02 | |
| 1 | 10 | 0.03 | 0.03 | |
| 1 | 11 | 0.01 | 0.02 | |
| 1 | 12 | -0.02 | -0.03 | |
| 1 | 13 | 0.04 | 0.04 | |
| 1 | 14 | -0.05 | -0.05 | |
| 1 | 15 | 0.02 | 0.02 | |
| 1 | 16 | 0.07 | 0.07 | |
| 1 | 17 | -0.03 | -0.03 | |
| 1 | 18 | 0.00 | 0.01 | |
| 1 | 19 | -0.02 | -0.03 | |
| 1 | 20 | 0.04 | 0.04 | |
| 1 | 21 | -0.01 | -0.02 | |
| 1 | 22 | -0.01 | -0.02 | |
| 1 | 23 | 0.03 | 0.03 | |
| 1 | 24 | 0.04 | 0.04 | |
| 1 | 25 | -0.03 | -0.04 | |
| 1 | 26 | -0.03 | -0.04 | |
| 1 | 27 | -0.03 | -0.03 | |
| 1 | 28 | -0.02 | -0.03 | |
| 1 | 29 | 0.01 | 0.02 | |
| 1 | 30 | -0.04 | -0.04 | |
| 1 | 31 | 0.07 | 0.08 | *FLAG |
| 1 | 32 | -0.05 | -0.05 | |
| 1 | 33 | -0.02 | -0.03 | |
| 1 | 34 | 0.01 | 0.02 | |
| 4 | 35 | -0.02 | -0.07 | |
| 4 | 36 | -0.09 | -0.09 | |
| 4 | 37 | 0.22 | 0.22 | |
| 4 | 38 | 0.00 | 0.03 | |
| 4 | 39 | -0.16 | -0.16 | |

In this example, for the item flagged the UDIF value exceeds +/- .075. That means item 31 seems to be showing DIF. That's the first stage.

Go back into the data file, and modify that sequence of "switches" you placed on the second line of the data file. For each item with a UDIF value greater than .075, a '0' should correspond to that column. Think of it in this way: by putting a '0' in a given column, you are effectively "switching off" that item from being included in the computation of the criterion score, thereby making the total test score a better matching criterion. From stage 1 we saw that item 31 seems to be DIF. Here's what the revised first few lines of the data files will look like (pay particular attention to the second line):

```
            11111111111111111111111111111144444
            11111111111111111111111111011111111
            00000000000000000000000000000000000
5050000474  11010110111011100101000010110011011313
5050000502  00100001000110000101011010101001000000
5050000577WM01011001001001000000000000000000012000
9002038068WM11111101011111111100101111100111003243
9002038228WF01000001001001000001100000000000011010
9002038710HF00001111110000000000100000000100112013
9002038827WF10001011110011101110001111111011122213
9002038836WM11111000100100011100010100110111031033
9002039000WF11111101101110111001110111110110142022
```

**Note #1**: Always make certain to align your datafile with the name of the datafile in your command file. If you change a datafile name in modifying the "switch" line or anything, be certain to change the name of the data file in the command file as well, or STDIF might read in a data set different than the one that you want.

After modifying the data file, just rerun 'stdif *filename*.cmd' at the DOS prompt to get new DIF indices based on the new criterion scores.

You may or may not see a substantial difference in the items being flagged for DIF between the first and second stages. That's O.K, either way.

**Note #2**: As you are running STDIF, take the time to look at the *filename*.LOG file produced by the program every so often. It provides a nice check for you to make sure the program is doing what you want it to do. In the .log file your command file is printed, as well as the line of maximum item scores and switches you put at the top of the data file. It also lets you know which examinees do not have a proper group identifier for the analysis you're trying to do and the number of examinees in the reference and focal groups you've specified at each of the score levels.

**Note #3:** Keep track (with handwritten notes or otherwise) of which items you are switching off in each analysis, and also the number of items that are "off".

## Part III: Summarizing Your Results

Item classification is clearly a critical part of these analyses. For each test, note how many items fall into each of three categories after Stage 2. Some rules of thumb for flagging items based on DIF statistics are given below.

| DIF statistic exceeding ±0.1 | DIF statistic between 0.075 and 0.1 (or -0.075 and –0.1) | DIF statistics between -0.075 and 0.075 |
|---|---|---|
| These are items which you will be looking at closer in order to try and infer sources of DIF | These items are flagged as potential DIF but are not studied for causes of DIF | Items not flagged as favoring one group or another |

**Your reports**: There are several tables and figures that you may want to produce for your reports.

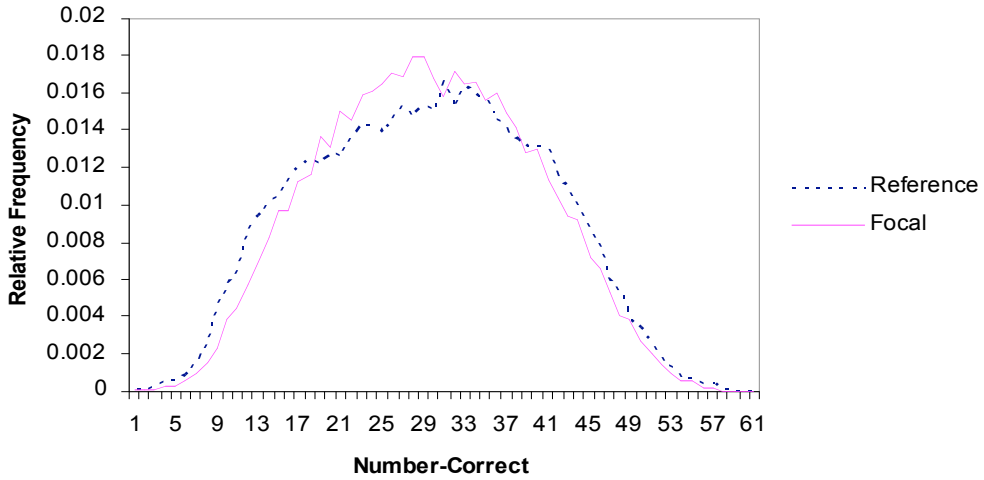
1) **Descriptive statistics of the test scores in each group** (mean, SD, N, coefficient alpha)

Table 1. Descriptive Statistics: Grade 4 Test
(Number of Items = 42; Maximum Score =72)

| Subgroup | N | Mean | SD | Reliability |
|---|---|---|---|---|
| Reference* | 38223 | 44.31 | 11.09 | .8945 |
| Focal* | 36339 | 47.48 | 10.79 | .8895 |
| Total | 74844 | 45.83 | 11.07 | .8927 |

*Insert the names of the reference and focal groups you are comparing here (Males and Females, Whites and Blacks, or Whites and Hispanics).

2) **Graph of total score distributions**: As illustrated below, this is a graph that the relative frequency of reference and focal group examinees at each score point.



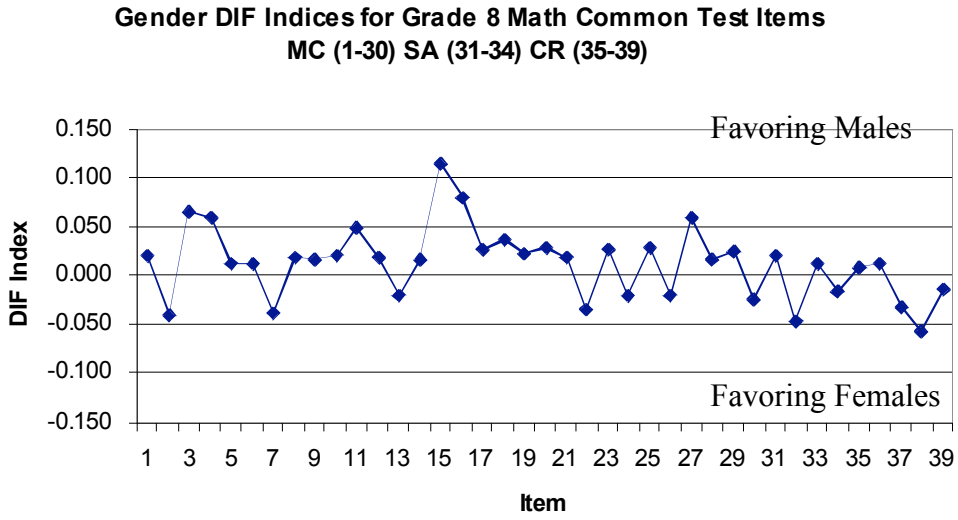
3) **Listing of the SDIF and UDIF statistics for stages 1 and 2** (for all items on test)

Table 2. Summary of DIF Indices[1]: Grade 4 English Language Arts

| Item | First Stage | Second Stage |
|---|---|---|
| 1 | .022 | .027 |
| 2 | -.005 | -.002 |
| 3 | .022 | .030 |
| … | … | … |
| … | … | … |
| 41 | -.027 | -.023 |
| 42 | -.029 | -.025 |

[1]Items were flagged at the 0.075 level.

4) **Presentation of the complete set of DIF indices**: This is a graph that visually represents the UDIF values for each item on the test.




5) **Summary of DIF Item Statistics**: This is a table, as shown below.

Table 3. Summary of DIF Item Statistics: Grade 4 Test (Male-Female)

| Number of Items Favoring Males | Number of Items Favoring Females | Number of Items | | |
|---|---|---|---|---|
| | | DIF < \|.075\| | \|.075\| to \|.10\| | DIF > \|.10\| |
| 32 | 11 | 6 | 3 | 0 |

6) **Presentation of the complete set of DIF plots**: Conditional p-value plots (i.e., p-values conditioned on total test score) are to be computed for each item. Both female and male results will be included in each plot. Three examples of conditional p-value plots are found on page 2 and 3 of this handout.

7) **Table mapping data and test questions**

| Item Number | Test Question Number[1] |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| 5 | 6 |
| … | … |
| … | … |
| 39 | 28 (ORC[2]) |
| 40 | 37 (ORC) |
| 41 | WP1[3] |
| 42 | WP2 |

[1] Item number refers to the item number as it appeared in the data file. This number was used throughout the DIF anlaysis. Test question number refers to the actual question number that appeared in the test booklet.
[2] ORC refers to Open Response Question.
[3] WP refers to Writing Prompt.

## References

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp.35-66). Hillsdale, NJ: Lawrence Erlbaum.

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

Robin, F. (2001). STDIF: Standardization-DIF analysis program [Computer program]. Amherst, MA: University of Massachusetts, School of Education.