ITEM RESPONSE THEORY LIKELIHOOD-RATIO TEST (IRT-LR) PERFORMANCE FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING (DIF) IN DIFFERENT SAMPLE SIZES AND DIFFERENT LEVELS OF (DIF) ITEMS

Aljoudeh Majed Mahmoud
University of Tabuk, SudiaArabia ,Tabuk
majed_jodeh@hotmail.com

ABSTRACT

The aim of this study is to examine the performance of the Item response theory Likelihood-Ratio test (IRT-LR) for detecting (DIF) in different sample sizes, and different levels of (DIF) items. For this purpose, Wingen3 software was used to generate four different sample sizes (250, 500, 750, and 1000), which represent responses on a 40 binary items test in two cases of (DIF) items. In the first case, some items were forced to be uniform DIF items in different levels of DIF (30% of all items), and non-uniform DIF items in the second case. The performance ratio of the IRT-LR method was investigated for detecting the DIF items in each case. The study concluded that at the sample size of (1000), the method showed a high percentage of performance in detecting the uniform DIF items of all levels, while the method performance decreased in its ability to detect the non-uniform DIF for all levels at all sample sizes that were dealt within the study.

Keywords: *Differential item functioning, Item response theory, Likelihood-ratio test, Levels of (DIF) items*

Introduction

In educational and psychological studies, it is difficult for researchers to verify the validity and reliability of the measurement; this is due to the complex nature of the variables that they deal with in their studies, as they express human characteristics. It is not easy to provide accurate measurement tools to measure such variables; therefore, it is difficult to control and adjust measurement errors, or even to know their type, value and direction.

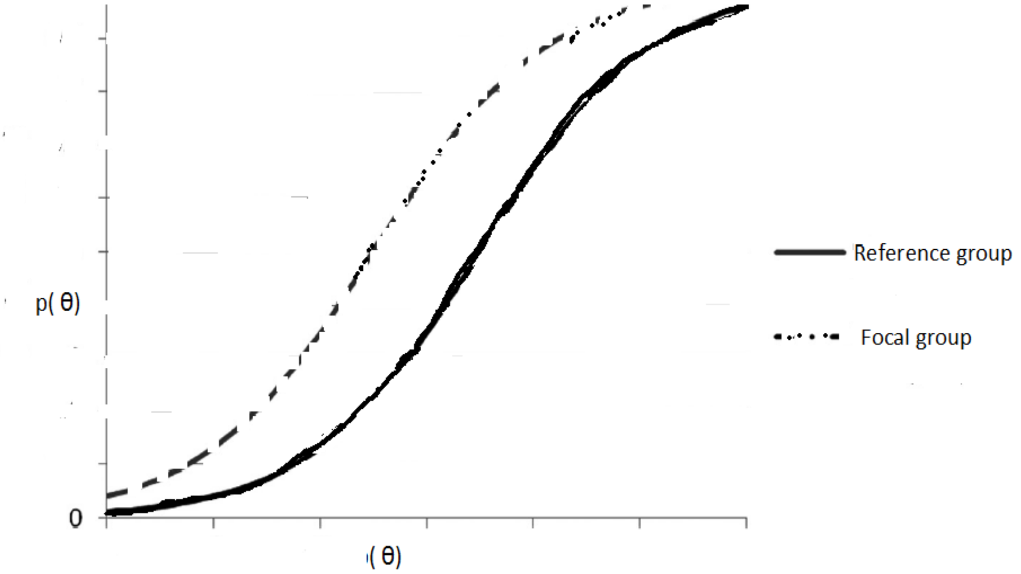
One factor affecting negatively on the validity and reliability of the measurement is the items bias, the presence of biased items reduces the reliability of the measurement tools, and the discussions related to them. The item's bias is usually the result of measurement errors, which generally affect the validity and reliability of the measurement. In the validity analysis procedures, it is important to search among the biased items for those that show differential item functioning (DIF), which can be detected through statistical methods.

Ellis and Raju (2003) explained that the DIF items are due to the presence of the biased items, and in the early 80's the term DIF began to take its place among researchers as an alternative to the term biased items. An item has DIF when its statistical psychometric properties vary for the groups that are matched on the attribute measured by the items. In the Item Response Theory framework, the DIF is defined as a difference in the conditional probabilities of responding to an item correctly in two or more groups. (Hidalgo and et al., 2004)

Some researchers have clarified that in addition to the main objective of DIF in detecting unfair items, it is necessary to use DIF analysis as a step in verifying the construct validity of the scale. (Walker & Beretvas, 2001).

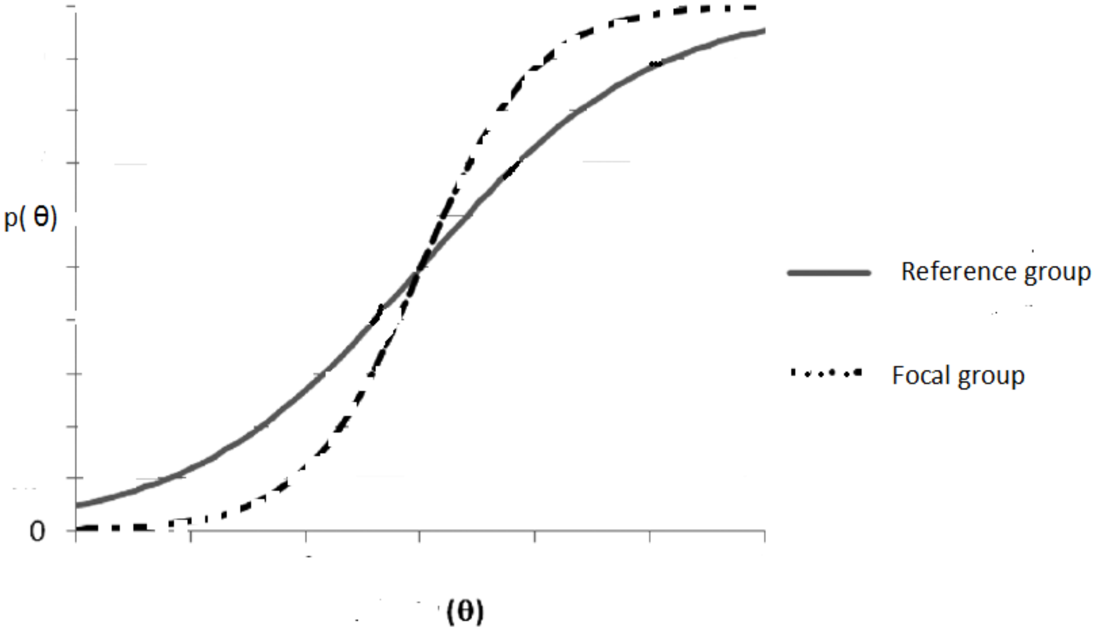
DIF analyses are most often conducted on two different groups, which are referred to as the reference and focal group. The reference group is typically the group that one hypothesizes may have an unfair advantage of obtaining the correct answer to a particular item. When this hypothesis is substantiated then an item is said to be functioning differentially against the focal group, or in favor of the reference group.(Walker, 2011). The analysis of the differential performance of the test items is an important aspect in order to ensure that the scores of the students have no bias in the test and that the theoretical construction of the measured trait is the same for all individuals who respond to the test. It is worth mentioning that the DIF of the test items can be uniform or non-uniform. Therefore, we can clarify the item characteristic curve (ICC) through item response theory (IRT). In the case of uniform DIF, the difference between the ICC in the reference and focal groups remains constant in all individuals' ability levels, as shown in Figure 1.

Figure 1: Uniform DIF



If the difference in the item characteristics curve differs between the reference and focal groups due to the different levels of ability, then the non-uniform DIF appears as shown in Figure 2.

Figure 2: Non-uniform DIF



It is difficult to detect the non-uniform DIF, so most DIF detection procedures designed for uniform DIF are unable to detect the non-uniform DIF (Walker, 2011).

Numerous DIF screening procedures have been presented in the literature. Most of these are based on theoretical foundations and the results of empirical comparisons studies. Some of these methods provide comparisons of DIF performance on a studied item after matching examinees on the ability of interest (Clauser & Mazor, 1998).

IRT Methods

IRT methods are not a single method, because they depend on the different models in the theory, and all of them are based on the principle of the difference for the estimated parameters of the item between the focal group and the reference group. A difference in the estimation of the difficulty parameter between the two groups leads to appearance of the uniform DIF, which I explained previously in Figure 1, and in the case of using the two parameter model (item difficulty, and item discrimination), the difference in the estimated values of the two parameters between the focal group and the reference group causes the non-uniform DIF as it is shown in Figure 2.

One of the common approaches, which researchers preferred is likelihood- Ratio test (IRT-LR); this method is based on calculating the LR value, which is given by the following equation:

$$\boldsymbol{Lc - (-2\, log\, log\, La)} \\ \boldsymbol{= -2\, log\, log\, Lc + 2\, log\, log\, La}$$

Where (Lc) is the log likelihood ratio of the compact model, the case in which no group differences are assumed to be present, (La) is the log likelihood ratio of the augmented model, the case in which one or more items are tested for DIF (Thissen et al., 1986; Cohen et al., 1996).

The (LR) values follow a chi-squared distribution with degrees of freedom equal to the number of parameters to be estimated in case that we have two groups (focal and reference groups). The significance means that there is a difference between the estimated parameters in the two models, subsequent tests are needed to compare the fit of the two models, with all item parameters except one (either A or B parameter) held the same. The main concept of IRT-LR is to assess whether or not the parameters for the item significantly differ across groups (Price, 2014).

In order to obtain accurate estimates for the parameters of the item in IRT, some researchers have indicated that the theory requires a large sample size (around 1000), and this depends on accurate estimates of the parameters of ability (Hambelton, 1989).

It is clear that there is an effect of the sample size in estimating the parameters of the item and the ability in IRT. When referring to the theoretical literature, it was found that there were many studies conducted in this regard, and reached somewhat different results. For example, the study of (Goldman & Raju, 1986) suggested that the minimum sample size for estimating accurate parameters in the (1PLM) model is (250), the (Guyer& Thompson, 2011) study is (300) and the (Thissen&wainer , 1982) study is about (500).

Some studies revealed that the more complex the model used, the greater the difference in the sample size used in their studies (stone, 1992; Weiss & Minden, 2012; Harwell &Janosky, 1991; Sahin& Anil, 2016; Yoes, 1995)

By reviewing the studies regarding sample size and its impact on the estimates of the parameters IRT, it is possible that a change in sample size will affect DIF as well.

Similar to the results of studies of the impact of sample size on the accuracy of parameter estimation in IRT, the results of the DIF studies were different and varied as well. In this regard, the results of (Gao, 2019) study, which compared six DIF detection Methods, indicated that all methods work better when large sample sizes are used as well the study of (Dorans, & Holland 1993) he suggested that whenever feasible, the largest possible sample size for both focal and reference groups should be used in DIF. While the (Jamali, J. et al., 2017) revealed that, the MIMIC method was recommended for detection of uniform-DIF when the focal group sample size is small.

The results of (Lee &Bulut& Suh, 2016) indicated that the ability of the MIMIC model to detect uniform DIF is higher than that of non-uniform. The study also found that when the length of the test and the sample size increases, the effectiveness of the method in detecting DIF increases.

The study of (Arikan&Ugural&Atar, 2016) aims to investigate the similarities and differences in four methods to detect DIF: MIMIC, SIBTEST, LR Logistic Regression, and Mantel-Haenszel MH. Different levels of sample sizes: (300, 600, 1000, 1200, and 2000) were used. The results of the study concluded that some items showed DIF in some methods but not all of them. For example, items (12, 13) did not show DIF in all methods on samples (300, 600), and on sample (1000) none of the methods showed DIF, while item (19) showed a common DIF in all methods on sample (1200), and on sample (2000) items (2, 3, 4) showed a common DIF in all methods. This study concluded that in a sample size of (2000) or higher that is more effective.

The likelihood- Ratio test (IRT-LR) method is based on IRT models. There are varying results of studies related to the influence of sample sizes on the accuracy of estimating parameters, as well as the results of DIF deducting studies, so the issue of selecting an appropriate sample size remains under consideration. This study comes through a simulation of different levels of sample size to investigate its effect on deducting DIF in different levels of DIF items. Reaching the appropriate sample size in a common method is scientifically useful and has an impact on the validity of the results of studies related to it.

WinGen3 software used to create four different sample sizes: (250, 500, 750, and 1000), which represent responses on a 40-item test of the binary items in two cases of DIF:

First case (Uniform DIF)

Twelve items were selected as to be uniform DIF items (30%), which are even-numbered items. Three levels of DIF are selected based on the difference in the item difficulty (B) parameter.

The items in which the difficulty difference is about (0.5) between the reference and focal groups are considered in the first level of the DIF (Low DIF), and the items in which the difficulty difference is about (1) are considered in the second level (Medium DIF) and the items with a difficulty difference of (1.5) are considered in the third level (High DIF).

Table No. 1 shows the distribution of items with the Pre-Uniform DIF in the three levels for each sample size.

Table #1: Distribution of Uniform DIF's items in the generated data.

Level of DIF	**Items Number**
Low DIF items	2, 4, 6, 8
Medium DIF items	10, 12, 14, 16
High DIF items	18, 20, 22, 24
Total / %	**12 / 30%**

Second case (Non-Uniform DIF)

As in the first case, twelve items (30%) were selected as Non-uniform DIF items, which are odd-numbered items. Three levels of DIF are selected based on the difference in the item (B) and (A) parameters. The items in which the parameter (A), (B) difference is about (0.5) between the reference and focal groups are considered in the first level of the DIF (Low DIF), and the items with (1) difference are considered in the second level (Medium DIF) and the items with difference of (1.5) are considered at the third level (High DIF).

Table No. 2 shows the distribution of items with the Pre-Non uniform DIF in the three levels for each sample size.

Table # 2: Distribution of Non-uniform DIF's items in the generated data.

Level of DIF	**Items Number**
Low DIF items	1, 3, 5, 7
Medium DIF items	9, 11, 13, 15
High DIF items	17, 19, 21, 23
Total / %	**12 / 30%**

The Data were analyzed using BILOG-MG software to conduct the uniform DIF items using 1PL -IRT model for generating data in case one and 2PL-IRT model for case two.

Results and Discussion

The current study aimed to investigate performance of the Likelihood-Ratio test for deducting DIF items in different sample sizes and levels of DIF items.

Table No. 3 shows the results of applying IRT-LR for data in case one. The table shows the item numbers that the method was able to reveal among those that contained a pre-DIF at the three levels of DIF in different sample sizes, as well as the item numbers that were revealed by the method and not selected as DIF items when generating data(Non pre-DIF items).

In addition, table No. 4 shows the differences between focal and reference groups in parameter "B" for these items.

Table # 3: Numbers of uniform DIF items after applying IRT-LR in case one.

Sample size	**Low DIF**	**Medium DIF**	**High DIF**	**Non pre- DIF items**
250	No DIF deducted	12	No DIF deducted	9, 27
500	No DIF deducted	No DIF deducted	No DIF deducted	No DIF deducted
750	No DIF deducted	No DIF deducted	No DIF deducted	3, 31,33,37
1000	6,8	10,14,16	18,20,22,24	11,13,15,17,21,26, 27,35

Table #4: the differences between focal and reference groups for DIF items in case one.

Sample size	Item Number	Absolute Group Difference	Standard Error	Difference/S.E
250	12	0.723	0.34	2.13*
	9	0.596	0.203	2.94*
	27	0.499	0.235	2.12*
750	3	0.787	0.083	9.48*
	31	0.276	0.081	3.41*
	33	1.627	0.102	15.95*
1000	6	0.392	0.104	3.77*
	8	0.273	0.11	2.48*
	10	0.527	0.12	4.39*
	14	0.232	0.107	2.17*
	16	0.206	0.104	1.98*
	18	0.528	0.105	5.03*
	20	0.414	0.115	3.60*
	22	0.343	0.108	3.18*
	24	0.21	0.106	1.98*
	11	0.929	0.128	7.26*
	13	0.808	0.111	7.28*
	15	0.705	0.128	5.51*
	17	0.853	0.11	7.75*
	21	1.252	0.132	9.48*
	26	0.23	0.108	2.13*
	27	0.476	0.123	3.87*
	35	0.367	0.104	3.53*

(*) Significant at 0.05

By referring to Table No.1, and Table No.3, we can summarize the percentage of performance of IRT-LR method in its ability to deduct the Uniform DIF in Table No. 5

Table # 5: performance percentage for deducting Uniform DIF using IRT-LR method.

Sample size	Low DIF	Medium DIF	High DIF	All
250	0%	25%	0%	8.3%
500	0%	0%	0%	0%
750	0%	0%	0%	0%
1000	50%	75%	100%	75%

By observing the results in Tables 3 and 5, it becomes clear to us that the method was not successful in detecting the DIF items in the medium sample sizes and the difference in the percentage of performance can be noted when the sample size of (1000) is used.

We conclude that the method will be more effective, when the large sample size (1000 and more) is used, and highly DIF items were found. Despite the ability of the method to reveal item (1) of the medium DIF level in the case of using a small sample size, it is difficult to rely on it, only 25% of the items with medium DIF were revealed. The percentage of what the method reached did not exceed 8.3% of the differential items at all levels.

These results may support researchers who have indicated that IRT theory requires a large sample size of (around 1000), and this depends on accurate estimates of the parameters of ability (Hambelton, 1989).

On the side of DIF studies with sample size, despite the different methods used, the result of this study also supports those studies that recommended the use of large sample sizes to obtain accurate results, (Dorans, & Holland 1993; Jamali, J. et al., 2017; Lee &Bulut& Suh, 2016; Arikan&Ugural&Atar, 2016).

We note From Table No. 3, that IRT-LR method revealed to us the presence of some DIF items, although they were not considered as such at the time of data generation (Non pre-DIF items), and this is suitable to be a research issue to interpret, but this may be due to the re-estimation of the parameters of the items and individuals when analyzing the responses according to the insertion of DIF items in the data file at the time it was generated.

Table No.6 shows the results of applying IRT-LR for data in case two. The table shows the item numbers that the method was able to reveal among those that contained a pre-DIF at the three levels of DIF in different sample sizes, and the Non pre-DIF items. In addition, table No.7 Shows the differences between focal and reference groups in parameter "B" for these items.

Table #6: Numbers of Non-uniform DIF items after applying IRT-LR in case two.

Sample size	**Low DIF**	**Medium DIF**	**High DIF**	**Non pre- DIF items**
250	No DIF deducted	11	No DIF deducted	No DIF deducted
500	1, 7	9, 11	No DIF deducted	29, 31
750	No DIF deducted	No DIF deducted	No DIF deducted	37
1000	No DIF deducted	11	21	36

Table #7: the differences between focal and reference groups for DIF items in case two.

Sample size	**Item Number**	**Group Difference**	**Standard Error**	**Difference/S.E**
250	11	0.404	0.134	3.01*
500	1	0.629	0.048	13.10*
	7	0.737	0.109	6.76*
	9	0.38	0.088	4.32*
	11	0.386	0.09	4.29*
	31	0.684	0.235	2.91*
750	37	0.23	0.108	2.13*
1000	11	0.666	0.069	9.65*
	21	0.865	0.334	2.59*
	36	0.557	0.217	2.57*

(*) Significant at 0.05

The more complex the model used in IRT, the more important the need for an appropriate sample size. Just as we noticed in previous studies that there were clear differences in their results about the appropriate sample size to use (stone, 1992; Weiss & Minden, 2012; Harwell &Janosky, 1991; Sahin & Anil, 2016; Yoes, 1995), we can notice from table no.4 that the performance of the method was not the same as it was in the first case.

We can summarize the percentage of performance of IRT-LR method in its ability to deduct the Non-Uniform DIF in Table No.8

Table # 8: performance percentage for deducting Non-uniform DIF using IRT-LR method.

Sample size	**Low DIF**	**Medium DIF**	**High DIF**	**All**
250	0%	25%	0%	8.3%
500	50%	50%	0%	33.3%
750	0%	0%	0%	0%
1000	0%	25%	%25	16.7%

The table No.8 shows, the ability of the method in general decreased for detecting Non-uniform DIF items and the highest performance percentage appeared when the average sample size was close to (500), reaching (33.3%) at all. I think this is a confusing result, and it points to a problem in the method procedures for non-uniform DIF detection. It is possible that we need to increase the sample size by more than (1000) for more accuracy. This result supports what Walker indicated, that it is very difficult to detect non-uniform DIF items, and therefore most of the methods designed to detect uniform DIF only, (Walker, 2011).

Each method has its own characteristics, problems, and it differs from one to another in its ability to detect the DIF items. The result of this study and other studies confirm the using of large sample size to reach more accuracy, so this study recommends using a sample size greater than (1000), it also recommends the studying of other factors that may be more important in this issue, such as the length of the test, the type of items, and others.

References

1. Arikan, C. A., & Uğurlu S. , & Atar B. (2016). A Dif and Bias Study by using Mimic, Sibtest, Logistic Regression and Mantel-Haenszel Methods..Journal of Education. 31(1): 34-52.
2. Clauser, B., &Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, 17(1), 31-44. doi:10.1111/j.1745-3992.1998.tb00619.x
3. Cohen, A.S., Kim, S.H.,&Wollack, J.A.(1996).An Investigation of the Likelihood Ratio Test For Detection of Differential Item Functioning .Applied Psychological Measurement, 20(1),15-26. doi:10.1177/014662169602000102
4. Dorans, N, Holland, P.(1993).DIF Detection and Description: Mantel-Hanszel and Standardization in p. Holland & H. Wamer(Eds).Differential item functioning (pp. 35-66) Hillsade, NJ: Lawrence Erlbaum Associates, Inc.
5. Gao, X.(2019). A Comparison of Six DIF Detection Methods. Master's thesis, University of Connecticut Graduate School,1411, https://opencommons.uconn.edu/gs_thesis/1411.
6. Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters:An empirical study. Educational and Psychological Measurement, 46(1), 11–21. http://dx.doi.org/10.1177/0013164486461002
7. Guyer, R., & Thompson, N. A. (2011). User's manual for Xcalibre 4.1. St. Paul, MN: Assessment Systems Corporation.
8. Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn(Ed.), Educational measurement (3rd ed., pp. 147–200). New York, NY: Macmillan.
9. Harwell, M. R., &Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement, 15(3), 279–291. http://dx.doi.org/10.1177 /014662169101500308
10. Hidalgo, M. D., &López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. Educational and Psychological Measurement, 64(6), 903–915. https://doi.org/10.1177/0013164403261769
11. Jamali, J ,Ayatollahi, S. &Jafari, P.(2017). The Effect of Small Sample Size on Measurement Equivalence of Psychometric Questionnaires in MIMIC Model: A Simulation Study. Bio Med Research International. Volume 2017, Article ID 7596101, https://doi.org/10.1155/2017/7596101
12. Lee, S., Bulut, O., and Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. Educational Psychol0gical Measurement. https://doi.org/10.1177/0013164416651116
13. Price, E.A. (2014).Item discrimination, model-data fit, and type I error rates in DIF detection using lord's chi 2, the likelihood ratio test, and the mantel-Hansel procedure (OrderNo.3671542). Available from ProQuest Dissertations &Theses Global.(1647205011). Retrieved from https://ezproxy.lib.uconn.edu/login?url=https://search.proquest.com/docview/1647205011?accountid=14518
14. Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. Educational Sciences: Theory & Practice,17,321–335. http://dx.doi.org/10.12738/estp.2017.1.0270
15. Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of Multilog. Applied Psychological Measurement,16(1),1 -16
16. Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika,47(4),397–412. http://dx.doi.org/10.1007/BF02293705
17. Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99(1), 118–128. https://doi.org/10.1037/0033-2909.99.1.118
18. Walker, C. M.(2011). What's the DIF? Why differential item functioning analyses

are an important part of instrument development and validation. Journal of Psychoeducational Assessment, 29(4), 364-376.

19. Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. Journal of Educational Measurement, 38, 147-163
20. Weiss, D. J., & Minden, S. V. (2012). A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG. St. Paul, MN: Assessment Systems Corporation.
21. Yoes, M. (1995). An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model. Saint Paul, MN: Assessment Systems Corporation.