International J. Soc. Sci. & Education
2016 Vol.6 Issue 1, ISSN: 2223-4934 E and 2227-393X Print

Effect of DIF Magnitudes, Focal Group Sample Size, and DIF Ratio on the Performance of SIBTEST

By

Sevilay Kilmen
Faculty of Education, Abant Izzet Baysal University, Gölköy, Bolu, TURKEY.

Abstract

This study focused on determining the effects of differential item functioning (DIF) magnitude, focal group sample size, and DIF ratio in a test of the statistical power and Type I control rate of the Simultaneous Item Bias Test procedure on simulated data. The percentages of DIF items were 10%, 20%, and 30% of the items on a 20-item test. Moderate and large DIF magnitudes were tested. Abilities were generated randomly from a normal distribution N(0,1) and changed for each replication. To examine the effect of sample size ratio on SIBTEST performance, 1000 examinees were generated for reference groups and varied numbers of examinees were generated for focal groups (1000, 500, and 250). 3600 1-0 data matrices were generated by WinGEN2. SIBTEST was used to calculate the DIF statistics. Power and Type I error rates were computed for each experimental condition, based on 100 replications. According to the results, conditions with moderate DIF had better Type I error control rates than those with large DIF. The power of SIBTEST increased as the focal group sample size increased. Maximum power (100%) was observed in DIF conditions that contained a 10% large DIF ratio.

Keywords: *Item Response Theory; differential item functioning, SIBTEST, DIF magnitude, DIF ratio, sample size ratio.*

1. Introduction

Differential item functioning (DIF) is "a manifestation of bias observed when examinees from different groups have different probability or likelihood of answering an item correctly, after controlling for ability" (Awuor, 2008). DIF occurs when individuals from different groups have unequal expected item scores or matching on the primary trait, attribute, or ability the test is designed to measure (Lopez Rivas, 2012). The two categories of DIF are uniform and nonuniform. Uniform DIF occurs when the probability of responding correctly to an item is uniformly higher for one of two groups across all levels of ability; these groups are called the reference and focal groups. Nonuniform DIF refers to the case where an item discriminates differently among the groups (Finch, 2005). Items that display DIF are a risk for test validity; these items must be identified and removed from the test in a process called item purification.
There are several DIF detection methods commonly used: SIBTEST (Shealy & Stout, 1993), Crossing SIBTEST (Li & Stout, 1996), Mantel-Haenszel (MH) Chi-square (Holland & Thayer, 1988), IRT likelihood ratio test (Thissen, Steinberg, & Wainer, 1993), logistic regression (Swaminathan & Rogers, 1990), multiple indicators, multiple causes (MIMIC) confirmatory factor analysis (MacIntosh & Hashim, 2003; Muthen, Kao, & Burstein, 1991), and MULTISIB (Stout, Li, Nandakumar & Bolt, 1997). SIBTEST is a non-parametric DIF detection method that does not require model calibration (Ackerman & Evans, 1992). SIBTEST can simultaneously evaluate DIF in several test items and allows one to select a matching subtest. For examinees, SIBTEST compares the average proportion correct on the subtest for the reference and focal group examinees (Narayanan & Swaminatthan, 1994). SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that were allocated to bins based on their scores in a "matching subtest". The statistical hypothesis tested by SIBTEST is (Stout & Roussos, 1996):

$$H_0: \beta_{UNI} = 0 \qquad H_1 = \beta_{UNI} \neq 0$$

Where β_{UNI} is the parameter specifying the amount of DIF for an item, and β_{UNI} is defined as the weighted expected mean difference in the probability of a correct response on an item between reference and focal group examinees of identical ability (Stout & Roussos, 1996).
Several studies examined the performance of SIBTEST under varying conditions. Narayanan & Swaminathan (1994) took the ability difference between the focal and reference groups, DIF ratio, and sample size as manipulated factors and they compared SIBTEST to the MH. They determined that SIBTEST and the MH procedure were equally powerful for detecting uniform DIF for equal ability distribution, but SIBTEST outperformed MH when the reference and focal group ability distributions were unequal. MH and SIBTEST are, to some extent, dependent on sample size; determine the power of these procedures is recommended for small samples, considering the ratio of the reference and focal group sample sizes. SIBTEST performance is affected by sample size and test length (Rogers & Swaminathan 1993; Roussos & Stout, 1996), ability differences, the number of items in the test, sample size ratios, and the percentage and magnitude of DIF (Awour, 2008; Atalay Kabasakal, Arsan, Gök, & Kelecioğlu, 2014); SIBTEST performance is not affected by mean differences in ability (Pei & Lie, 2010). While SIBTEST is affected by the percentage of DIF, it can adequately detect DIF, even when 60% of the items contained DIF and sample sizes were at least 1,000 examinees per group (Gierl, Gotzmann, & Boughton, 2004). Lei and Li (2013) have examined the effects of small sample sizes and DIF magnitudes on SIBTEST and other DIF detection methods; they determined that larger sample sizes positively influence SIBTEST performance.

This study focused on determining SIBTEST performance of SIBTEST under varied conditions. After examining the literature, the above-mentioned studies did not examine DIF ratios, DIF magnitudes, and sample size ratios in a fully-crossed design, which allows us to examine the interactions of variables and DIF performance. The goal of this study was to investigate the effects of DIF magnitude, focal group sample size, and DIF ratio on SIBTEST's statistical power and Type I control rate.

2. Method

Manipulated factors
In this study, simulated test data were used to compare 18 different test conditions, varying the DIF magnitude, focal group sample size, and DIF ratios. In simulation studies, the DIF items were known a priori; therefore, Type I error and power were estimated from the percentages of false positives and true positives apart from replications. A Type I error occurs when an item indentified as DIF, but DIF was not simulated. The percentage of detections of items simulated to be DIF was used an empirical estimate of the power (Lee, Cohen & Toro, 2009). The simulation study to investigate the power and Type I error rates of SIBTEST with manipulated factors was conducted as follows:

- Sample size of focal group: 250, 500, and 1000;
- Number of items with DIF: 2 (10%), 4 (20%), and 6 (30%) items in the 20-item test length
- Magnitude of DIF: 0.4 per DIF item and 0.8 per DIF item

Data Generation
A computer program called WinGen (Han, 2006) was developed to generate dichotomous and polytomous item response data for several IRT models and for many conditions that arise in practice. WinGen provides dialog input to introduce DIF or item parameter drift in simulated data. With multiple file read-in options in WinGen, a user can have multiple examinee groups and multiple sets of items/tests (Han & Hambleton, 2007). In this study, twenty one-dimensional, dichotomously-scored item tests were generated by WinGen2 according to the two parameter logistic IRT model. The a parameters ranged from 0.533 to 1.399, with mean 0.954 and standard deviation 0.196. The b parameters were generated randomly from a normal distribution. In the first step of data generation, one simulative test form, called the reference group test form, and six test forms were generated with DIF for focal groups. According to Shepard, Camilli, and Williams (2008), differences in the b parameter of 0.20 were the least detectible, 0.35 produced moderate DIF, and 0.64 produced large DIF. These cut points were used in this study. The

DIF type was modeled as uniform. These test forms consisted of 20 items each and were created in terms of DIF magnitude and ratios:

- test form with two moderate DIF items
- test form with four moderate DIF items
- test form with six moderate DIF items
- test form with two large DIF items
- test form with four large DIF items
- test form with six large DIF items

Table 1: Item Parameter Values for Generating Simulated Data under the 2PL Model

Items	a Parameters	Reference Group b Parameters	Focal Group b Parameters					
			%10		%20		%30	
			Moderate DIF	Large DIF	Moderate DIF	Large DIF	Moderate DIF	Large DIF
1	0.940	-0.586						
2	0.700	0.562						
3	1.050	0.889						
4	1.071	-1.446						
5	1.159	1.295						
6	1.399	0.172						
7	0.786	0.589						
8	0.957	-1.298						
9	1.225	-1.511						
10	1.037	0.144						
11	0.698	0.872						
12	1.032	0.867						
13	1.064	-0.137						
14	0.934	-2.167						
15	0.879	-1.449					-1.049	-0.249
16	0.933	0.419					0.819	1.619
17	0.836	0.938			1.338	2.138	1.338	2.138
18	0.983	0.333			0.733	1.533	0.733	1.533
19	0.856	-1.254	-0.854	-0.054	-0.854	-0.054	-0.854	-0.054
20	0.533	1.245	1.645	2.445	1.645	2.445	1.645	2.445

Uniform DIF was simulated by keeping the a parameters for the reference and focal groups the same but varying the b parameters for the two groups. As shown in Table 1, two (items 19 and 20), four (items 17, 18, 19, and 20), and six items (items 15, 16, 17, 18, 19, and 20) were modeled to display uniform DIF of moderate magnitudes by increasing the focal group's b parameters by 0.4. Items were modeled again to display large uniform DIF by increasing the focal group's b parameters by 0.8. The item parameters for the focal group were the same as those for the reference group except items whose parameters were manipulated to show medium and large uniform DIF.

In the second step, abilities were generated randomly from a normal distribution N(0,1) via WinGen2. These person ability values were changed for each replication. To examine the effect of sample size ratio on SIBTEST performance, 1000 examinees were generated for reference groups and varied numbers of examinees were generated for focal groups: 1000, 500, and 250. The combinations between the reference

and focal groups were done in the ratios of 1:1, 1:2, and 1:4; in each condition, 100 replications were conducted. Unique data sets were generated for both the reference and focal groups in each replication. Thus response data of 3600 items were generated (1800 for focal groups and 1800 for reference groups). DIF values were obtained using SIBTEST.

Evaluation criteria

Type I error was defined the proportion of times a non-DIF item was flagged incorrectly as a DIF item across replications; in other words, the number of false positives divided by the number of replications (Lopez Rivas, 2012). In a simulation study, a Type I error occurs when an item indentified as DIF, but DIF was not simulated (Lee, Cohen & Toro, 2009). When the Type I error rate is high, it means that non-DIF items are incorrectly flagged as DIF items. On the other hand, power was defined as the number of times an item known to exhibit DIF is flagged by a DIF detection method; therefore, it is the number of true positives divided by the number of replications (Lopez Rivas, 2012). The percentage of detections of items simulated to be DIF was used an empirical estimate of the power (Lee, Cohen & Toro, 2009). When the power is high, it means DIF items are correctly identified. The statistical power and Type I error rates were computed over 100 replications.

3. Findings and Discussion

The results are presented in Table 2. Table 2 contains the power and Type I error rates as a function of the magnitude of DIF (moderate and large), DIF ratios in the tests (10%, 20%, and 30%), and focal group sample size (250, 500, and 1000).

Table 2: ***Type I Error and Statistical Power Results of SIBTEST***

Focal Group Sample size*	DIF magnitude	DIF ratio	Type I error rate	Power
1000	Moderate	10%	0.041	0.795
		20%	0.115	0.790
		30%	0.204	0.678
	Large	10%	0.077	1
		20%	0.291	0.990
		30%	0.316	0.899
500	Moderate	10%	0.006	0.720
		20%	0.108	0.789
		30%	0.097	0.613
	Large	10%	0.089	1
		20%	0.185	0.990
		30%	0.188	0.806
250	Moderate	10%	0.027	0.640
		20%	0.134	0.614
		30%	0.220	0.373
	Large	10%	0.096	1
		20%	0.110	0.760
		30%	0.180	0.473

*Reference group sample size: 1000

The SIBTEST results with the smallest Type I error rate had 10% moderate DIF for all sample sizes. For all conditions in which the percentage of DIF items was 30%, the Type I error rate was greatly inflated. For the aforesaid conditions, the condition with the highest Type I errors was the 30% large DIF

condition a focal group sample size of 1000. When focal groups sample sizes were 500 with moderate 10% DIF, there were fewer Type I errors than in the other conditions. The type I error rates increased as the proportion of items showing DIF increased (from 10% to 30%), which concurs with the results of Narayanan and Swaminathan (1994).

When current research findings are compared with Awuor (2008)'s and Atalay Kabasakal et al.'s (2014) findings, the power rates of these studies were higher than the current results, possibly due to different moderate and large DIF magnitudes, a parameters, b parameters, and test lengths in Awuor's (2008), Gierl, Gotzmann, & Boughton's (2004) and Atalay Kabasakal et al.'s (2014) studies. For example, 20-item tests were used in this study, while 50-item tests were used in Awuor's (2008) study. In Gierl, Gotzmann, & Boughton's (2004) study, the proportion of DIF was manipulated to be 20, 40, and 60 % of the 40 item test. Also, focal and reference group sample size were equal. Higher power results were obtained in this study possibly due to different test conditions when they were compared with current study.

DIF magnitude affected SIBTEST performance in this study; the power of SIBTEST increased as the DIF magnitudes increased from 0.4 to 0.8. Large DIF magnitude conditions produced large statistical power. SIBTEST had the most sufficient power under 10% large DIF conditions. The results support the findings of Awour (2008) and Narayanan and Swaminathan (1994).

Low power for DIF detection was obtained when the focal group sample size was small. The poorest power was noticed when focal group sample sizes were 250. In this research, the power of SIBTEST increased as focal group sample size increased. In other words, sample size positively influences SIBTEST performance. These results support the findings of Awour (2008), Finch (2005), Lei and Li (2013), Gierl, Gotzmann, and Broughton (2004), González-Romá et al. (2006), Narayanan and Swaminathan (1994), Rogers and Swaminathan (1993) and Zheng, Gierl & Cui (2007). According to Narayanan & Swaminathan (1994), this result was not surprising because empirical distributions are expected to approach theoretical distributions as sample size increases. According to Gierl, Gotzmann, and Broughton (2004), SIBTEST provided adequate DIF detection because incorrect item rejections were less than 5% and correct rejections were greater than 80% when DIF was balanced and the sample sizes were at least 1000 examinees per group. According to González-Romá et al. (2006), a sufficient power threshold should be 70%. With large DIF and equal sample sizes (N=1000), SIBTEST procedures showed sufficient statistical power; however, moderate DIF conditions produced insufficient power (≤70%). Moreover, it was determined that the power of SIBTEST increased as the DIF ratio decreased. In contrast, in Atalay Kabasakal et al.'s (2014) study, sample size ratio and DIF ratio did not affect the power of SIBTEST; this may be due to the different test lengths and different sample size ratios in Atalay Kabasakal et al.'s (2014) study.

4. Limitations and Recommendations

This study had some limitations. The first limitation was that only focal group sample sizes, DIF magnitudes, and DIF ratios were manipulated in this study; in future research, different variables can be manipulated. DIF type was modeled as uniform in the current study; the effects of nonuniform DIF were not estimated. Another limitation was that abilities were modeled with a two-parameter logistic model and normal distributions. Additionally, the focal and reference groups had similar ability differences. According to Finch (2005) and Wang and Yeh (2003), the difference in abilities between the reference and focal groups affected the Type I error; therefore, the ability difference between the focal and reference groups and different Item Response Theory models (one parameter, three parameters, and multidimensional models) can be researched. The performance of SIBTEST can also be examined against real data where the grading is 1-0 and multiple categories are carried out together or graded as multiple categories with different sample sizes. This research was limited to the SIBTEST method. Type I

errors and power can be researched under different conditions with different DIF detection methods.

References

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement, 18* (4) 329–342.

Atalay-Kabasakal, K., Arsan, N., Gök, B. & Kelecioğlu, H. (2014). Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory & Practice, 14*(6), 2186-2193. doi: 10.12738/estp.2014.6.2165

Awour, R. A. (2008). Effect of unequal sample sizes on the power of DIF detection: An IRT- based Monte Carlo Study with SIBTEST and Mantel-Heanszel procdures. Unpublished doctoral dissertation, Virginia Polytechnic Institude and State University, USA.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29* (4), 278-295.

Gierl, M. J.,& Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1,* 249-270.

Gierl, M. J., Gotzmann, A., & Boyghton, K. A. (2004). Performance of SIBTEST when the percent of DIF items is large. *Applied Measurement in Education, 17*(3), 241-264.

Han, K.T. & Hambleton, R. K. (2007). User's Manual for WinGen: Windows Software that Generates IRT Model Parameters and Item Responses1Center for Educational Assessment Research Report No. 642. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Holland, P. W.,& Thayer, D. T. (1987). Notes on the use of the log-linear models for fitting discrete probability distributions (ETS Research Rep. No. RR–87–31). Princeton, NJ: Educational Testing Service.

Lei, P. W. & Li, H. (2013). Small-sample DIF estimation using SIBTEST, Cochran's *Z*, and Log-Linear smoothing. *Applied Psychological Measurement, 37* (5), 397-416. doi:10.1177/0146621613478150.

Li, H.-H.& Stout, W. (1996).A new procedure for detection of crossing Differential Item Functioning. *Psychometrika, 61,* 647-677.

Lopez Rivas, G. E. (2012). Detection and Classification of DIF Types Using Parametric and Nonparametric Methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and Logistic Regression Procedures. Unpublished doctoral dissertation, University of South Florida, USA.

Narayanan, P. & Swaminathan H. (1994). Performance of the Mantel-Haenszel Simultaneous Item Bias Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement, 18* (4), 315-328.

Pei, L. K., & Li, J. (2010). Effects of unequal ability variance on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement, 34* (6), 453-456. doi: 10.1177/0146621610367789.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17* (2), 105-116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DTF as well as item bias/DIF. *Psychometrika, 59*, 159-194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985).Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22* (2), 77-105.

Stout,W., Li, H. -H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*(3), 195–213.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual*. Urbana: University of Illinois.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27* (4), 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Weiner, H. (1993). Detection of differential functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Wang, W.-C. & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27* (6), 479-498. doi: 10.1177/0146621603259902.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007, April). *Using real data to compare DIF detection and effect size measures among mantel-Haenszel, SIBTEST and Logistic Regression procedures.* A paper presented at the annual meeting of the National Council on Measurement in Education: Chicago, ILL.

Bolt, D. M. (2000). A SIBTEST Approach to testing DIF hypotheses using experimentally designed test ıtems. *Journal of Educational Measurement, 37*(4), 307–327. doi: 10.1111/j.1745-3984.2000.tb01089.x

Lee, Y. S., Cohen, A. & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review, 10* (3), 365-375 doi:10.1007/s12564-009-9039-7