TPM Vol. 33, No. 1, 2026
ISSN: 1972-6325
https://www.tpmap.org/

EVALUATING THE EFFECTIVENESS OF THE SIMULTANEOUS ITEM BIAS TEST IN DETECTING DIFFERENTIAL ITEM FUNCTIONING ACCORDING TO GUESSING PARAMETER VARIATION AND SAMPLE SIZES

AMER ALMARABHEH
DEPARTMENT OF FAMILY AND COMMUNITY MEDICINE, COLLEGE OF MEDICINE AND HEALTH SCIENCES, ARABIAN GULF UNIVERSITY, MANAMA, BAHRAIN, ORCID NO: 0000-0002-4279-2371, E-MAIL: amerjka@agu.edu.bh.

MOHAMMAD ALQUDAH
DEPARTMENT OF PSYCHOLOGY, COLLEGE OF EDUCATIONAL SCIENCES, TAFILA TECHNICAL UNIVERSITY, JORDAN, E-MAIL: mohqud@ttu.edu.jo.

ABSTRACT
The Simultaneous Item Bias Test (SIBTEST) compares the performance of different groups on individual test items, considering their overall ability levels. It provides valuable insights into the potential sources of bias in the assessment and aids in making informed decisions regarding item inclusion or modification to enhance the fairness and validity of the test. The current study aims to assess guessing parameters and sample size on the effectiveness or performance of Statistics of the Simultaneous Item Bias (SIBTEST) as a method to detect differential item functioning (DIF) according to a 3-Parameter Logistic Model (3PLM) using Item Response Theory (IRT). The study adopted the experimental methodology, which controls the variables. The study participants are individuals with ability levels (θ) randomly generated from a standard normal distribution N (0,1). WinGen-3 program was used to generate data, and SIBTEST software was used for calculating the statistic (β) value. To achieve the study objectives, data have been generated for a (50) item test for both the reference and the focal groups. Three levels of guessing parameters (0.10, 0.15, and 0.20) were chosen for the focal group, whereas the guessing parameter of the reference group was stabilized at (0.20). Three sample sizes were used (250, 500, and 1000) individuals. To judge the effectiveness of statistics (β), both type one error and Test power were employed as measurements. The findings revealed that the test power ratio increases when the difference in the value of the guessing parameter decreases between the two groups with the increase in sample size. Moreover, the type one error was within the defined significance level ($\alpha < 0.05$) when the value of the guessing parameter was approximated between the two groups, with an increase in sample size.
Keywords: Guessing Parameter, Simultaneous Item Bias, Differential Item Functioning, Logarithmic Model, Simulation study.

INTRODUCTION

Contemporary measurement scientists have been urged to undertake innovative research efforts seeking the development of a new Psychometric theory that could be used to solve urgent problems. A modern theory for measurement emerged, known as Item Response Theory (IRT), which is based on varied maxims. Hence, many measurement tools and educational and psychological measurements are being reviewed based on the IRT to measure the participants' performance (Hambleton & Swaminathan, 2013), (Raju et al., 1995).
Differential item functioning (DIF) occurs when items exhibit a differential probability of a correct response between two groups after statistically controlling for ability. Hence, items with DIF could be described as biased items in one of the groups with reasons not related to individuals' abilities (Hou et al., 2014). DIF is one of the most significant aspects of test building, as it threatens test building validity and reliability, as well as parameter estimates of ability and items. That could ruin the results and make it difficult to make educational decisions in light of invalid data (Diaz et al., 2021). To distinguish between DIF, bias, and test fairness, we need to clarify that test fairness is achieved whenever the performance of participants with the same ability level is equal in test items, regardless of their culture, gender, or race. The test item is biased whenever individuals with the same ability have unequal chances of responding correctly to an item. Bias could exist whenever the difference in performance is

not true among two groups; however, DIF is a necessary condition but not enough to reach bias. Some items may have DIF but are not biased whenever the performance difference is due to real differences in individuals' abilities (Almarabheh & Alshammari, 2020; Finch, 2005; Pei & Li, 2010).

The Simultaneous Bias Test is a statistical test designed to simultaneously detect DIF present in one or more items of a test (Lei & Li, 2013). The SIBTEST approach tackles the DIF issue from a multidimensionality standpoint, utilizing a multidimensional nonparametric IRT model (Shealy & Stout, 1993), which is considered the first method based on Item Response Theory (IRT) to detect DIF in multiple items simultaneously is known to have limited power in identifying nonuniform DIF (Sandilands et al., 2013). The SIBTEST measures latent ability based on observed scores on valid items, comparing reference and focal groups based on correct answers for suspected items. The procedure assesses the presence of a statistically significant distinction between the probability of correctly responding to an item by the reference and focal groups. Using a matching technique based on the examinees' total test results, SIBTEST computes a weighted mean difference between the two groups. When a significant difference is found, it means that an item exhibits DIF. Iteratively, the SIBTEST operates until every questionable item is eliminated from the valid subset. The matching criterion is applied to the final subgroups of DIF-free items (Clauser & Mazor, 1998), (Jiang & Stout, 1998), (Weese et al., 2022).

Several research studies have highlighted the importance of various factors in ensuring accurate DIF detection, whether based on Classical Test Theory or Item Response Theory models (DeMars & Wise, 2010; Finch & French, 2014; Finch & French, 2007; Kilmen, 2016; Kolen & Brennan, 2004; Magis et al., 2010). These factors include the choice of model, test length, sample size, DIF item ratio, pseudo-guessing parameter, ability distribution, corresponding degree, DIF effect, and parameter estimation conditions. In the current study, the focus was narrowed down to two key factors believed to have the greatest impact on the accuracy of DIF detection, namely the difference in pseudo-guessing parameters and the sample size. However, despite considering these factors, questions remain regarding the factors that influence the accuracy and effectiveness of DIF detection methods. Therefore, the primary objective of this study is to assess the performance of the SIBTEST method in the accuracy of detection of differential item functioning in terms of type one error and test power, considering different guessing parameters and sample sizes within the framework of the three-parameter logistic model.

Objectives and Questions of the Study

This study aims to assess the performance of the SIBTEST method in detecting Differential Item Functioning (DIF). Thus, the following research questions are examined:

1) Does the accuracy of detecting the differential item functioning of the SIBTEST method differ according to the guessing parameter variation?
2) Does the accuracy of detecting the differential item functioning of the SIBTEST method differ according to the sample size variation?
3) Does the accuracy of detecting the differential item functioning of the SIBTEST method differ according to the interaction between the variation of the guessing parameter and sample size?

RESEARCH METHOD

Study Design

Based on a hypothetical assessment, random samples were simulated under known and controlled population settings in the present study, which utilized the Monte Carlo simulation study method.

Participants

The study participants are individuals with ability level (θ) randomly generated from a standard normal distribution N (0, 1). All unidirectional tests with dichotomous responses (1: correct, 0: incorrect) have the same parameter distribution as specified in the study. The sample sizes used in the study were 250, 500, and 1000 individuals.

Data Generation and study process

WinGen-3 is a computer application designed to provide item response data for both polytomous and dichotomous items (Han, 2007), considering various conditions and Item Response Theory models. The study focused on comparing item responses between two groups: the focal group and the reference group. In this study, a set of (50) dichotomously scored items was generated using WinGen-3, following the 3-parameter logistic model (3PLM). The characteristics of the test items were specified to be similar for both groups, assuming a normal distribution of individual ability parameters and item difficulty parameters with a mean of (0) and a standard deviation of (1). Additionally, three levels of pseudo parameters were chosen for the focal group (0.10, 0.15, 0.20), and three sample sizes (250, 500, 1000) were considered. To investigate type one error, the parameters of the two groups were set to be equal, ensuring that the items were non-differential according to the 3PLM. Item response data, consisting of (50) items without DIF, were simulated, except for items 49 and 50, which exhibited DIF. These simulated examinee responses were used to calculate examinee scores and generate theta coefficients for different sample sizes and levels of the guessing parameter. Subsets of items were then selected from the simulated data to achieve the desired values of coefficient alpha. **Table 1** presents the number of items selected to simulate population parameters, including difficulty, discrimination, and guessing, for the 48 non-differential items used to study type one error.

Table 1. Population item parameters estimated for data simulation.

Item	Disc.	Difficulty	Guessing	Item	Disc	Difficulty	Guessing
1	0.390	-1.578	0.126	25	0.377	0.792	0.138
2	0.489	1.674	0.119	26	0.334	-1.958	0.142
3	0.400	-0.256	0.125	27	0.484	-2.004	0.148
4	0.372	-2.169	0.103	28	0.441	-0.789	0.108
5	0.341	0.186	0.126	29	0.427	-0.173	0.130
6	0.419	2.720	0.127	30	0.357	0.355	0.144
7	0.446	-0.193	0.127	31	0.467	0.195	0.129
8	0.430	-0.768	0.140	32	0.408	0.210	0.142
9	0.394	0.696	0.119	33	0.420	-0.030	0.114
10	0.457	-0.771	0.100	34	0.433	0.574	0.147
11	0.452	0.510	0.123	35	0.347	0.565	0.104
12	0.398	0.205	0.129	36	0.438	-0.224	0.130
13	0.322	0.074	0.134	37	0.318	0.488	0.145
14	0.328	-1.164	0.133	38	0.459	1.386	0.143
15	0.472	-1.090	0.129	39	0.448	0.913	0.139
16	0.356	0.402	0.102	40	0.430	0.489	0.116
17	0.441	1.015	0.106	41	0.470	0.520	0.150
18	0.434	-0.211	0.124	42	0.477	-0.607	0.148
19	0.400	-0.230	0.117	43	0.335	-0.432	0.104
20	0.406	-0.367	0.106	44	0.323	-0.639	0.113
21	0.327	-2.333	0.107	45	0.426	-0.749	0.125
22	0.413	-1.146	0.116	46	0.333	-1.142	0.112
23	0.419	-1.411	0.130	47	0.349	1.677	0.111
24	0.456	-0.55	0.100	48	0.381	0.212	0.123

To study statistical test power, two differential test items were generated within a 50-item test. The two DIF items were designed such that the difficulty parameters differed between the reference and focal groups, while the discrimination parameter (0.90) remained the same for both groups. Specifically, the difference in difficulty parameters between the reference and focal groups was set to 1, representing a specific magnitude of DIF. The guessing parameters were held constant (0.20) across the two DIF items for both groups. The parameters for the two DIF items are presented in **Table 2**.

Table 2. The parameter values for the two DIF items simulated for both the reference and focal groups

DIF Item	Reference Group			Focal Group		
	Discrimination	Difficulty	Guessing	Discrimination	Difficulty	Guessing
49	0.90	-0.50	0.20	0.90	0.50	0.20
50	0.90	0.50	0.20	0.90	1.5	0.20

DIF was assessed using SIBTEST software (Weese, 2022). To ensure reliable results, the data were replicated 100 times for each condition, following guidelines from (Zwick, 2012), and (Penfield et al., 2009), resulting in 100 data matrices for each case. The same procedures were applied in detecting proper differential items and were divided into 100 groups, and the percentage of the number of times for each case was considered evidence for the power of the statistical test.

The statistical test power was categorized based on Cohen's guidelines, as adapted by , into three levels for practical interpretation. The classification criteria (DIF classification) were as follows: if the statistical test power is:

1. Less than 0.70, it is categorized as negligible.
2. Greater than or equal to 0.70, it is categorized as moderate power.
3. Greater than or equal to 0.80, it is categorized as a large power.

In the current study, a proper differential item was considered indicative of statistical test power if it achieved a value of 0.70 or more (moderate or large DIF levels). While these categories provide a framework for practical interpretation, the continuous values of statistical test power are also reported to retain the full detail of the results and provide a nuanced understanding of the findings.

Evaluation criteria

Two criteria were used to determine the quality of performance of the DIF procedures. The first is the type one error rate, which measures how many times a non-DIF item was incorrectly flagged as a DIF item across replications. It is calculated as the ratio of false positives to the total number of replications (Atalay Kabasakal et

al., 2014),(Lopez, 2012). In a simulation study, a type one error occurs when an item is identified as DIF, but DIF is not simulated (Lee et al., 2009). When the type one error rate is high (greater than the nominal alpha value of 0.05), it means that non-DIF items are incorrectly flagged as DIF items. The second criterion for evaluation is power, which can be described as the number of times a DIF-exhibiting item is reported by a DIF detection method. Essentially, it's the ratio of true positives to replications (Lopez, 2012). The percentage of detections of items simulated to be DIF was used as an empirical estimate of the power (Lee et al., 2009). When the power is high (more than or equal to 0.70) shows that the method's power is sufficient, which means DIF items are correctly identified. The statistical power and type one error rates were computed over 100 replications.

Data analysis

WinGen3 software (Han, 2007) was used to generate data based on the 3PL model. The Statistical Package for the Social Sciences (SPSS) version 28 was used to analyze the impact of simulation factors on the independent variables with all replications for each condition in each simulation study. To achieve a maximum standard error, each simulation study has 100 replications. The SIBTEST Version 1.7 software was used to obtain the statistical value beta (β) of the SIBTEST method (Weese, 2022), (Stout & Roussos, 1995). To estimate likelihood functions for θ and β, maximum-likelihood estimation (ML) techniques are used. An alpha level of 0.05 was applied for DIF detection.

RESULTS

The findings of this study are presented in three sections. The type one error rate is examined in the first section using a variety of parameters, including sample size and guessing parameter. The second section is to look into the test power rate and determine how often a DIF detection method (SIBTEST) flags an item that is known to display DIF. The third and final section presents the measurement of the impact of each guessing parameter and sample size on the two assessment criteria, which are the test power rate and type I error.

Type one error rate

To identify the effect of the study variables: guessing parameter and sample size on type one error. The data were analyzed by calculating the statistical value (β) of the SIBTEST method and calculating the type I error in each situation during the test. It is the number of cases that showed DIF of one item divided by several test items by the frequency of trials, then comparing the results with the inference level at ($\alpha < 0.05$). The results showed that type one error is statistically significant when the guessing parameter of the reference group is (0.15) the sample size is (500-1000), and type one error reached (0.044, 0.048), less than ($\alpha < 0.05$). Also, the results showed type one error is statistically significant when guessing the parameter of the reference group (0.20) and sample size (1000, 500), as type one error reached (0.038, 0,042) less than the specified significance level (**Table 3**).

Table 3. Type one error for SIBTEST according to the guessing parameter and sample size.

C. Parameter CF/CR	Sample size nF=nR	No. of items	Overall	SIBTEST Statistics (β)	
				Frequency	Percent
0.20/0.10	250	48	4800	359	0.075
	500			304	0.063
	1000			263	0.055
0.20/0.15	250	48	4800	289	0.060
	500			228	0.048
	1000			209	0.044
0.20/0.20	250	48	4800	244	0.051
	500			203	0.042
	1000			184	0.038

Test Power rate

To identify the effect of study variables: guessing parameters and sample size on the test power of the SIB Test, data from two items showing regular DIF were extracted. The two items have equal discrimination parameters in the reference and focal groups, whereas the two groups' difficulty parameters varied. Data was reextracted (100) times, and the number of times items showed DIF correctly was calculated and divided by the frequency. Test power was categorized using the amended (Cohen) measurement. The results indicated that the illustrated statistical significance of SIBTEST power was high in one cell when guessing the parameter of the focal group is (0.10), and the sample size is (1000). The results showed that the test power rate was on average in three cells: one of them when guessing the parameter of the focal group (0.10) and sample size (500), and in the other two cells when guessing the parameter of the focal group (0.15) and sample size (500,1000) (**Table 4**).

Table 4. The test power of the statistical (β) according to the guessing parameter and sample size.

C. Parameter CF/CR	Sample size nF=nR	No. of items	Overall	SIBTEST Statistics (β)	
				Frequency	Percent
0.20/0.10	250	2	200	133	0.665
	500			157	0.785
	1000			163	0.815
0.20/0.15	250	2	200	128	0.640
	500			141	0.705
	1000			148	0.740
0.20/0.20	250	2	200	119	0.595
	500			123	0.615
	1000			131	0.655

Effect of guessing parameter and sample size on the type I error and test power

The study extracted the (β) value of the SIBTEST, as well as type one error and test power, considering different guessing parameters. The Results (**Table 5**) revealed that the type I error rates for (β) were found to be lower than the significance level ($\alpha < 0.05$) when the guessing parameter of the focal group was set to 0.20. However, the results indicated that the type one error for (β) exceeded statistical significance when the guessing parameter of the focal group was 0.15 or 0.10. Regarding the results of (β) for different guessing parameters of the focal group, a test power rate of average degree was observed when the guessing parameter was set to 0.10. Furthermore, the results related to different sample sizes revealed that the type one error of (β) was lower than the statistical significance level ($\alpha < 0.05$) when the sample size was 1000. However, the type one error rate exceeded statistical significance for sample sizes of 250 and 500; also, the manuscript results indicated that the test power was average when the sample size was 500 or 1000. The results related to the interaction between the guessing parameter and sample sizes showed that type one error for (β) was statistically significant when the guessing parameter of the focal group (0.20, 0.15) and sample size (1000, 500) (**Figure 1**). As for the test power, the results showed that (β) power rate is an average degree when guessing the parameter of the focal group (0.10) and sample size (500), and when guessing the parameter of the focal group (0.15) and sample size (1000, 500) as well. For more details, see Figures (1) and (2), which show differences in type one error rates and (β) test power with different interactions between guessing parameters and sample size (**Figure 2**).

Table 5. Rate of Type One Error and Test Power Based on Variation in the Guessing Parameter and Sample Size.

Variables	Categories	Rate of type one error	Test Power
Guessing Parameter	0.10	0.064	0.755
	0.15	0.051	0.695
	0.20	0.044	0.622
sample size	250	0.062	0.633
	500	0.051	0.710
	1000	0.046	0.737

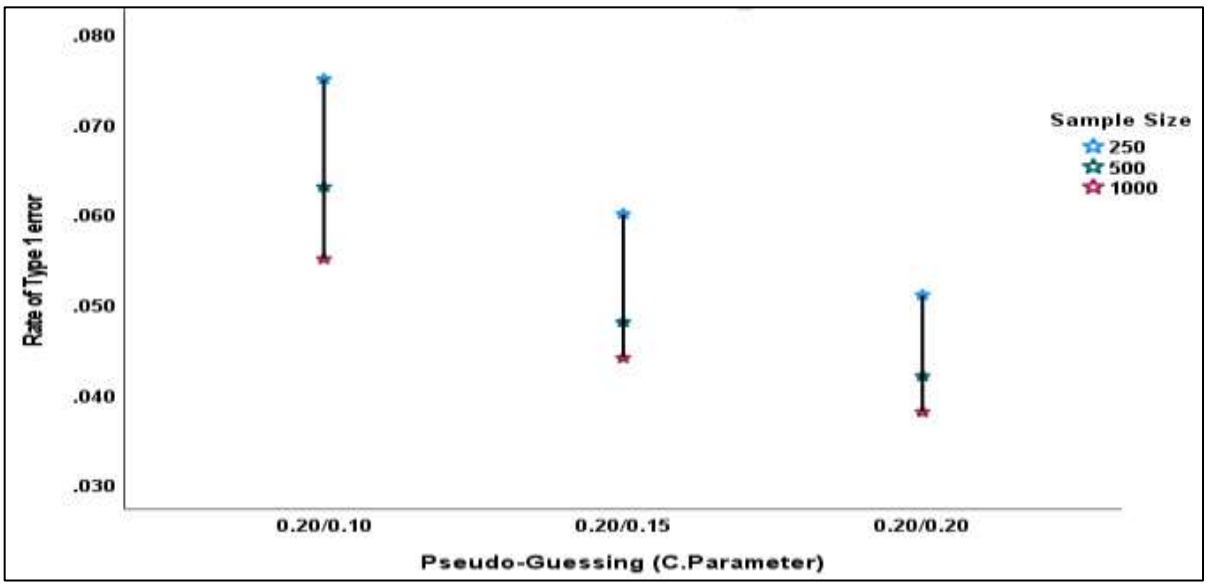

Figure 1. The rate of type one error based on Variation in the Guessing Parameter and Sample Size

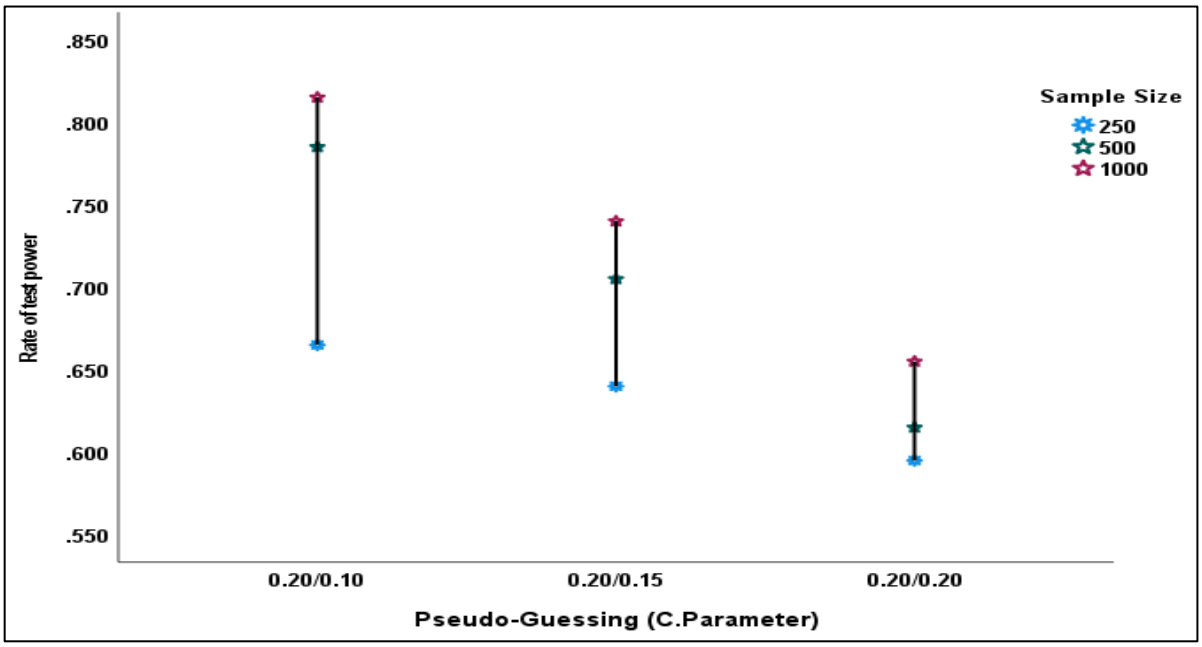

Figure 2. The rate of test power based on Variation in the Guessing Parameter and Sample Size

DISCUSSION

This study aimed to assess the impact of various pseudo-guessing parameters and sample sizes on the efficacy of the Simultaneous Item Bias Test according to two criteria: type one error and test power, using the Three-Parameter Logistic Model. DIF analysis plays a crucial role in ensuring fairness and validity during test development. Accurate and specific detection of DIF is vital not only for identifying its presence but also for determining the type of DIF, enabling measurement specialists and psychometricians to make necessary adjustments to enhance fairness. This study sought to assess the effectiveness of SIBTEST in accurately detecting DIF using the 3PLM and considering variations in both pseudo-guessing parameters and sample sizes.

The findings reveal that type one error is statistically significant when the guessing parameter for the focal group falls within the range of (0.20, 0.15), coupled with sample sizes of 1000 and 500. Additionally, SIBTEST demonstrates a high detection rate (0.70) when the guessing parameter for the focal group is set at 0.10, and the sample size is 1000. Furthermore, the study indicates that the power rate of SIBTEST increases with larger sample sizes, and type one error is statistically significant when the guessing parameter for the focal group is 0.20 (Finch & French, 2014).

The increase in Type I errors with higher guessing parameter values can be attributed to the role of the guessing parameter in the 3PLM. The guessing parameter introduces a non-zero probability of correct responses for low-ability individuals, effectively adding variability to item response patterns. When the guessing parameter differs significantly between the reference and focal groups, this variability may be misinterpreted as DIF, leading to inflated Type I error rates. Conversely, the higher power observed when c differs between groups may reflect the increased sensitivity of SIBTEST to systematic differences in item performance caused by true DIF. This highlights the dual impact of the guessing parameter as both a potential confounder and a contributor to DIF detection accuracy, particularly under conditions of large sample sizes where statistical power is inherently greater.

Moreover, the results underscore the power and effectiveness of SIBTEST in detecting DIF, particularly when large sample sizes and differing guessing parameters are employed for the reference and focal groups. The dissimilarities in item parameters (difficulty, discrimination, and guessing) between the reference and focal groups significantly contribute to the efficacy of DIF detection, particularly in the case of larger sample sizes (Finch & French, 2014). This is due to SIBTEST's reliance on correcting regression equating, which is contingent upon stability and, in turn, benefits from increased sample sizes.

These findings align with those of (Finch & French, 2007), (Finch & French, 2014), which examined the effects of different guessing parameters, ability distributions, and sample sizes on the accuracy of DIF detection using various statistical methods. Their results similarly demonstrated that Type I error is more likely to be statistically significant when guessing parameters are higher and equal between the reference and focal groups. Additionally, when ability distributions are matched and sample sizes are large, Type I error rates tend to stabilize, consistent with findings from (Weese et al., 2022), (Atalay Kabasakal et al., 2014), (Uysal et al., 2019).

The current study's results also align with those of (DeMars & Wise, 2010), who investigated how quick-guessing behavior among test-takers can contribute to DIF detection. Their findings suggest that certain instances of DIF detection may be attributable to differences in guessing behavior rather than genuine item bias, reinforcing the importance of carefully considering the role of the ccc parameter in DIF analysis.

Regarding test power, the findings are consistent with prior research by (Kolen & Brennan, 2004), (Weese et al., 2022), and (Çepni & Kelecioglu, 2021), which demonstrated that larger sample sizes enhance the power of SIBTEST, leading to more accurate identification of DIF. This is likely because larger sample sizes reduce standard errors, improve parameter estimation precision, and amplify the statistical sensitivity of SIBTEST to detect true DIF.

In summary, the study highlights the intricate interplay between the guessing parameter and sample size in determining the performance of SIBTEST. While higher guessing parameters can increase variability and potentially inflate Type I error, they also enhance the ability to detect true DIF when coupled with larger sample sizes. These findings underscore the importance of selecting appropriate test design parameters to optimize the balance between Type I error control and test power. Future research should explore a wider range of guessing parameters and their interactions with discrimination and difficulty parameters to further illuminate the mechanisms underlying DIF detection. Additionally, expanding the scope of simulation replications and incorporating alternative DIF detection methods could provide further validation and generalizability of these findings.

CONCLUSION

This study examined the effectiveness of the SIBTEST method in detecting differential item functioning (DIF) under varying conditions of sample size and guessing parameter. The findings demonstrated that the efficacy of DIF detection improves with larger sample sizes, while smaller sample sizes, particularly when coupled with high guessing ratios, decrease accuracy. A positive association was observed between the guessing parameter and Type I error, with higher guessing parameters leading to an increased likelihood of Type I error.

The study acknowledges certain limitations that may influence the generalizability of the findings. First, the relatively small sample sizes (250, 500, and 1000), though reflective of practical constraints in testing contexts, may have limited statistical power and precision. Future research should consider larger sample sizes to enhance robustness. Second, while the majority of test items were designed with low discrimination parameters, the two DIF items were intentionally given higher discrimination values to fulfill their intended purpose in the study and to examine DIF detection under conditions where the DIF items are more discriminative, as such items may exhibit more pronounced DIF effects. Future studies could use DIF items with discrimination values that are more consistent with the other test items to isolate the impact of DIF more effectively. Finally, the Monte Carlo simulations employed 100 replications to balance computational efficiency with study objectives; increasing the number of replications in future research could improve the precision of Type I error and power estimates.

Future investigations are also warranted to extend these findings. Researchers should explore alternative DIF detection methods, including both Classical Test Theory (CTT) and Item Response Theory (IRT)-based approaches, while considering variables such as sample size, item length, guessing parameters, and the type of logistic model applied. Additionally, future studies should focus on examining both uniform and nonuniform DIF under various conditions and employing multiple DIF detection methods to provide a more comprehensive understanding of DIF. It is also important to extend the analysis to polytomous data, which was beyond the scope of this study.

Despite these limitations, the current study provides valuable insights into the effectiveness of SIBTEST in detecting DIF. By addressing these limitations and exploring additional methods, DIF types, and variables, future research can build upon these findings to enhance our understanding and application of DIF detection methodologies.

ACKNOWLEDGEMENT: N.A

REFERENCES

1. Almarabheh, A. J., & Alshammari, S. R. (2020). Detection of Sex-Related Differential Item Functioning in Raven's Standard Progressive Matrices Test Using the Mantel-Haenszel Method.
2. Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioglu, 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 and Practice*, *14*(6), 2186-2193.
3. ÇEPNİ, Z., & KELECİOĞLU, H. (2021). Detecting differential item functioning using SIBTEST, MH, LR and IRT methods. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 267-285.
4. Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. *Educational Measurement: issues and practice*, *17*(1), 31-44.
5. DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, *10*(3), 207-229.
6. Diaz, D., Brooks, G., & Johanson, G. (2021). Detecting differential item functioning: Item Response Theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education*, *8*(2), 376-393.

7. 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.
8. Finch, W. H., & French, B. F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling*, *56*(1), 25.
9. Gotzmann, A., & Boughton, K. (2004). A comparison of type I error and power rates for the Mantel-Haenszel and SIBTEST procedures when the group differences are large and unbalanced. In: American Educational Research Association.
10. Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
11. Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459.
12. Finch, W., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and psychological Measurement*, *67*(4), 565-582.
13. Hou, L., la Torre, J. d., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.
14. Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291-322.
15. Kilmen, S. (2016). Effect of DIF magnitudes, focal group sample size, and DIF ratio on the performance of SIBTEST. *International Journal of Social Sciences and Education*, *6*(1), 91-98.
16. Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking.
17. 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.
18. 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.
19. Lopez, 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.
20. Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, *42*(3), 847-862.
21. Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement*, *34*(6), 453-456.
22. Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38-49.
23. Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353-368.
24. Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing*, *13*(2), 152-174.
25. Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.
26. Stout, W., & Roussos, L. (1995). SIBTEST users manual [Computer program manual]. *Urbana-Champaign: University of Illinois, Department of Statistics*.
27. Uysal, I., Ertuna, L., Ertaş, F. G., & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 133-148.
28. Weese, J. D. (2022). DIFSIB: A SIBTEST Package. *Applied Psychological Measurement*, *46*(1), 68-69.
29. Weese, J. D., Turner, R. C., Ames, A., Crawford, B., & Liang, X. (2022). Reevaluating the SIBTEST classification heuristics for dichotomous differential item functioning. *Educational and Psychological Measurement*, *82*(2), 307-329.
30. Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i-30.