Article

Educational and Psychological Measurement
73(3) 458–470
© The Author(s) 2012
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164412467033
epm.sagepub.com

Effect of Multiple Testing Adjustment in Differential Item Functioning Detection

Jihye Kim[1] and T. C. Oshima[1]

Abstract

In a typical differential item functioning (DIF) analysis, a significance test is conducted for each item. As a test consists of multiple items, such multiple testing may increase the possibility of making a Type I error at least once. The goal of this study was to investigate how to control a Type I error rate and power using adjustment procedures for multiple testing, which have been widely used in applied statistics. In the simulation, four distinct DIF methods were performed under various testing conditions. The methods were the Mantel–Haenszel (MH) method, the logistic regression (LR) procedure, the Differential Functioning Item and Test (DFIT) framework, and Lord's chi-square test. As an adjustment procedure, the Bonferroni correction, Holm's procedure, or the Benjamini and Hochberg (BH) false discovery rate was applied. The results showed the MH and the LR clearly benefited from Holm's and the BH adjustments, whereas the DFIT and Lord's chi-square test did not require adjustments for conditions under this study.

Keywords

multiple testing, DIF, Type I error rate

A differential item functioning (DIF) study often involves significance testing of multiple items (i.e., the whole test). Such multiple testing may increase the possibility of committing a Type I error at least once (Shaffer, 1995), which leads to a high possibility of incorrectly identifying non-DIF items as DIF items. Falsely identifying DIF items can weaken the validity of the assessment. Hence, the quality of a test assessment is related to a Type I error rate and how to control the inflation of its rate.

[1] Georgia State University, Atlanta, GA, USA

Corresponding Author:
Jihye Kim, Georgia State University, 30 Pryor Street, Atlanta, GA 30303-3083, USA.
Email: jkim59@gmail.com

Adjustment procedures in multiple testing can be effective in controlling a Type I error rate (Shaffer, 1995). Although several DIF researchers (e.g., Penfield, 2001; Thissen, Steinberg, & Kuang, 2002) have demonstrated the usefulness of adjustment procedures, such as the Bonferroni correction (Bonferroni, 1936) or the Benjamini–Hochberg (BH) false discovery rate (Benjamini & Hochberg, 1995), it appears to be the exception rather than the rule to apply any type of adjustment procedures in applied DIF literature. Thus, the comparison of adjustment procedures should be of interest.

This study sought an effective way of controlling a Type I error rate by applying three adjustment procedures of multiple testing that have been commonly used as statistical methods in social science. Adjustment procedures are often referred to as controlling for ‘‘fishing expeditions.’’ Three existing adjustment procedures were considered, the Bonferroni correction, the Holm’s (1979) procedure (also known as the improved Bonferroni’s procedure), and the BH false discovery rate, since these three procedures are simple and easy to implement.

The Bonferroni correction has been used as one of most common adjustments for several decades in statistics. Holm’s (1979) procedure is an improved procedure that is more powerful than Bonferroni’s procedure (Holland & Copenhaver, 1987). Holm’s procedure is similar to Bonferroni’s, but it is known to be less conservative because it is less corrective as the number of tests increases and is based on the ordered p-values from each test. The ordered p-value methods are strong for controlling a Type I error rate when the test statistics are independent (Shaffer, 1995). Whereas the Bonferroni correction and the Holm’s procedure seek to control a family-wise Type I error rate, the Benjamini and Hochberg (1995) procedure controls expected false positive discovery rates (FDR) by defining a sequential p-value procedure.

In an attempt to compare different types of DIF methods, four distinct DIF methods were selected: the Mantel–Haenszel (MH) method (Holland & Thayer, 1988), the logistic regression (LR) procedure (Swaminathan & Rogers, 1990), the Differential Functioning Item and Test (DFIT) framework (Raju, Linden, & Fleer, 1995), and Lord’s chi-square test (Lord, 1980). The first two methods are referred to here as non-IRT (non–item response theory) methods as they do not require item calibration using the IRT. In particular, as a non-IRT based method, the MH method and LR procedure were selected in this study, since those methods are easy to access and have been used as popular statistical methods in social science areas, including DIF research. Furthermore, they are distinct in their approach to significance testing.

For the non-IRT based methods, the MH method and LR have been used to identify DIF items for more than 20 years. The MH method is an extension of the traditional two-way chi-square test of independence to the situation in which three variables are completely crossed, namely, group membership, performance on the item, and any number of levels of the attribute (Fidalgo & Madeira, 2008; Holland & Thayer, 1988; Mantel & Haenszel, 1959). In MH, we assume that the odds of getting the item correct across all score levels are the same in both focal and reference groups. Then the hypothesis is tested whether there is a significant association of a common odds ratio between groups. LR is the procedure that is used widely in statistical literature, which uses a model that links a categorical outcome with one or more predictor variables that can be either continuous or categorical (Swaminathan & Rogers, 1990; Zumbo, 1999). It is a simple and traditional way for detecting DIF items, where the group variable and conditioning variable (total score) are predictors in the model. Then, corresponding chi-squared values of predictors are tested to see if there is any significant difference.

For the IRT-based methods, DIFT compares two item characteristic curves (ICCs), and Lord's chi-square test compares item parameter estimates. DFIT is similar to the area measures (Raju, 1988), but instead of measuring the area between two ICCs, it calculates the average squared difference between two ICCs. See Oshima and Morris (2008) for more details. The noncompensatory DIF (NCDIF) index in DFIT has a unique significance test that relies on computer simulation. Oshima, Raju, and Nanda (2006) proposed a simulation-based significance test for NCDIF in which the cutoff value for each item is determined by $(1 - \alpha)$ percentile rank score from a frequency distribution of NCDIF values under the no DIF conditions over many replications. More information on Lord's chi-square test can be found in Hambleton, Swaminathan, and Rogers (1991).

As for the two IRT-based methods, Lord's chi-square test is based on a theoretical chi-square distribution, whereas the significance test for NCDIF in DFIT is based on an empirical distribution resulting from computer simulation. Lord's chi-square test was chosen as a contrast technique to DFIT. DFIT is a newer and more versatile IRT-based DIF technique than the Lord's chi-square test, as it can be used with various data types (dichotomous/polytomous) in various IRT models (unidimensional/multidimensional).

Previous researchers have used adjustment procedures in DIF studies. The comparison of adjustment procedures between non-IRT based and IRT-based methods, however, has been rarely investigated. Since the DIF detection procedures from the non-IRT based and the IRT-based methods were quite different, they may react differently under different adjustment procedures. The main goal of this study is to investigate the effect of adjustment procedures for multiple testing in the context of DIF studies:

The three research questions in this study are as follows:

Research Question 1: What is the effect of adjustment procedures on the test-wide Type I error rate for the four DIF methods?

Research Question 2: What is the effect of adjustment procedures on the power for the four DIF methods?

Research Question 3: Which of the three adjustment procedures (the Bonferroni correction, Holm's procedure, and the BH false discovery rate) works the best for the four DIF methods?

Review of Related Literature

Three Adjustment Procedures

The Bonferroni correction and Holm's procedure are a widely used adjustment procedure of controlling multiple tests (Holland & Cohenhaver, 1987). In particular, Holm's procedure has been used often in the areas of clinical trials and biology (Soulakova, 2009). In the measurement area, Reschly, Busch, Betts, Deno, and Long (2009) used Holm's procedure for evaluating appropriate curriculum-based measurements of reading outcomes.

DIF analysis is based on multiple testing associated with testing every item one at a time, thus the use of adjustment procedures has been applied in assessing DIF among multiple groups and evaluating DIF items (Crane, van Belle, & Larson, 2004; Penfield, 2001; Stark, Chernyshenko, & Drasgow, 2006). Penfield (2001) applied the Bonferroni correction to the MH method, and the results showed the good control of a Type I error rate. Steinberg (2001) also applied the BH method on the IRTLR (item response theory log-likelihood ratio) procedure, for the comparison between the *p* values of the chi-squares and the BH-corrected *p* values. The result showed that the BH-corrected *p* values emerged slightly larger than observed *p* values. It could be interpreted that the BH procedure of correcting *p* values is more effective in reducing a Type I error than the observed *p* values.

Several research studies found that the BH false discovery rate was proved to yield much greater power than the widely used Bonferroni, which is based on controlling Type I error rates (Benjamini & Hochberg, 1995; Thissen et al., 2002; Williams, Jones, & Tukey, 1999). However, one finding shows that there is little difference among those adjustment procedures. In particular, with the presence of nonuniform DIF condition and ordinal items, four adjustment procedures (the Bonferroni, Holm, Hochberg, and Sidak procedures) showed the same results in ordinal logistic regression (Crane et al., 2004). See the appendix for an illustrated example of the three adjustment procedures.

Method

Study Design for the Simulation Study

Data generation. Using SAS, the dichotomous scored data were generated with a three-parameter IRT model. Examinee abilities were assumed to follow the standard normal distribution. The probability of a correct response to an item was calculated based on prespecified item parameters from Oshima et al. (2006; see Table 1). The basis probability was generated at random from the uniform distribution. When comparing the calculated probability and a basis probability, if the basis probability was less than the calculated probability, the simulated item response was scored as *correct* (1); otherwise, it was scored as *incorrect* (0).

Sample size and test length. The total sample size selected for this study was 2,000 (1,000 for the reference group and 1,000 for the focal group) and 1,000 (500

Table 1. Item Parameters for the 40-Item Test and the 20-Item Test.

Item		Reference		Focal (10%)	
40	20	*a*	*b*	*a*	*B*
1	1	0.55	0		
2		0.55	0		
3	2	0.73	−1.04		
4		0.73	−1.04		
5	3	0.73	0		
6		0.73	0		
7	4	0.73	0		
8		0.73	0		
9	5	0.73	1.04		
10		0.73	1.04		
11	6	1	−1.96		
12		1	−1.96		
13	7	1	−1.04		
14		1	−1.04		
15	8	1	−1.04		
16		1	−1.04		
17	9	1	0	1	0.3
18		1	0	1	0.3
19	10	1	0	1	0.5
20		1	0	1	0.5
21	11	1	0	1	0.7
22		1	0	1	0.7
23	12	1	0		
24		1	0		
25	13	1	1.04		
26		1	1.04		
27	14	1	1.04		
28		1	1.04		
29	15	1	1.96		
30		1	1.96		
31	16	1.36	−1.04		
32		1.36	−1.04		
33	17	1.36	0		
34		1.36	0		
35	18	1.36	0		
36		1.36	0		
37	19	1.36	1.04		
38		1.36	1.04		
39	20	1.8	0		
40		1.8	0		

for the reference group and 500 for the focal group). Sample size of 1,000 each would provide a sufficient sample size for item calibration for the IRT methods. Since the current recommendation of the DFIT is to use equal sample sizes, unequal sample sizes were not considered in this current study. For the test length, lengths of 20 items and 40 items were chosen as common assessments are constructed with equal or fewer than 40 items (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994; Raju et al., 1995; Rogers & Swaminathan, 1993; Roussos & Stout, 1996).

Percent of DIF level. Table 1 shows how DIF was embedded. The percentage of DIF was constant at 15%. All *a*-, *b*-, and *c*-item parameters were set to be the same for both reference and focal groups, except that three items (Item 9, 10, and 11) in the 20-item test and 6 items (Item 17, 18, 19, 20, 21, and 22) in the 40-item test where *a*-parameter and *c*-parameter for the reference and focal group were equal. The magnitude of DIF was determined by the *b*-parameter. Swaminathan and Rogers (1990) also used the *b*-difference for the focal and reference groups and defined .64 as the baseline for purporting large DIF (Category C). For medium DIF, they specified .48 as the *b*-difference. In this study, 0.7 difference of *b*-parameter was denoted as large DIF magnitude (i.e., Items 21 and 22 in the 40-item test, and Item 11 in the 20-item test), 0.5 difference of *b*-parameter as medium magnitude (i.e., Items 19 and 20 in the 40-item test, and Item 10 in the 20-item test), and 0.3 difference of *b*-parameter as small magnitude (i.e., Items 17 and 18 in the 40-item test, and Item 9 in the 20-item test).

DIF detection procedure. This study examined only uniform DIF across groups, whereas there are two types of DIF: uniform DIF and nonuniform DIF. For the DIF detection of non-IRT methods, ability matching for reference and focal groups was performed by calculating the total scores before detecting DIF items (Zwick, Donoghue, & Grima, 1993) and then the statistics were calculated using the MH method and the logistic regression. The DIF detection for IRT-based methods was conducted by two-stage linking procedures, and then the statistics were calculated using the DFIT method and Lord's chi-square test.

Results

The overall Type I error rate and degree of power were assessed when multiple items were tested concurrently, although individual tests of each item were conducted at a Type I error rate of .05. In order to evaluate the inflation of a Type I error rate, Bradley's (1978) liberal robustness criterion range of .025 to .075 was used.

The test-wide Type I error rate for this study was calculated as follows. The simulation conducted 100 replications (test sets). First, for each replication (a test set), the occurrences of false positives out of all non-DIF items were counted. Then, the proportion of these counts was calculated per test set, focusing on the practical point of view as to how many items were falsely identified as DIF items in each test set. A Type I error rate reported in this study is the average of these proportions over 100 replications.

The power in this study was also of interest. We examined the trend of power with three types of DIF magnitude (large, medium, and small). The power was calculated in the similar way as the Type I error rate was calculated. First, the proportion of true positives out of all DIF items with a specific DIF magnitude was calculated for each replication. Then the power with a corresponding DIF magnitude is the average of these proportions over 100 replications.

The Effect of the Three Adjustment Procedures

For non-IRT based methods, most unadjusted Type I error rates of the MH method and LR were greater than 0.10 (Table 2). For instance, before adjustments, MH's Type I error rate with a sample size of 1,000/1,000 for the 40-item test was 0.12, which means that about 4 out of 34 non-DIF items were flagged for DIF. When put it into a practical testing situation, this is problematic as almost half of the flagged items on a test are non-DIF items (assuming up to 6 DIF items are also flagged). The larger sample size as well as longer test resulted in more inflation of Type I error rate. However, it is noteworthy that increasing the test length from 20 to 40 resulted in only about 20% increase in the Type I error rate.

For the power, the main concern is the detection rates for medium and large DIF items, as those are the items typically to be considered for removal. Bonferroni's procedure often resulted in an undesirable decrease in power. For example, with the 20-item, 500/500 condition, MH's power decreased from .91 to .63 for large DIF. Although the Type I error rate decreased from .07 to .02, the advantage does not seem to justify the noticeable lack of power. On the other hand, both Holm's and BH's procedures maintained a good balance of decreasing the Type I error rate while keeping the power, namely, keeping the power for large DIF larger than .80. The results from MH and LR are similar, but the inflation of Type I error was slightly worse for LR. For IRT-based methods, interestingly, DFIT and LR both did not show any inflation of Type I error before adjustment. Therefore, the adjustment was not necessary. When the adjustments were made anyway, Lord's chi-square suffered substantial loss of power. On the other hand, DFIT maintained reasonable power, especially with an adequate sample size (1,000 in each group) along with medium to large DIF. The comparison of DIFT and Lord's chi-square clearly shows that DFIT is a preferred method over Lord's chi-square as the latter seriously lacks power especially with a smaller sample size.

In sum, the results show that MH and LR benefited from Holm's or BH's adjustment procedures at all test lengths and sample sizes considered in this study. The Type I error rates for those methods were reduced after adjustment procedures (mostly within Bradley's liberal criterion, between .025 and .075, while maintaining reasonable power). On the other hand, IRT-based procedures did not benefit from the adjustment procedures as the inflation of Type I errors was not observed under conditions in this study.

Table 2. Type I Error Rate and Power of Four DIF Methods.

	Condition		Type I Error Rate				Power Rate (With DIF Magnitude)			
Method	Test Length	Sample Size	Unadjusted	Bonferroni		BH	Unadjusted (in %)	Bonferroni (in %)	Holms (in %)	BH (in %)
MH	20	500/500	0.07	0.00	0.02	0.06	0 (S)	0 (S)	0 (S)	0 (S)
							64 (M)	24 (M)	50 (M)	62 (M)
							91 (L)	63 (L)	88 (L)	91 (L)
		1,000/1,000	0.10	0.01	0.03	0.09	42 (S)	9 (S)	27 (S)	40 (S)
							89 (M)	57 (M)	81 (M)	88 (M)
							100 (L)	98 (L)	100 (L)	100 (L)
	40	500/500	0.09	0.00	0.02	0.07	33 (S)	4.5 (S)	15.5 (S)	32.5 (S)
							77.5 (M)	25 (M)	55.5 (M)	76 (M)
							96.5 (L)	67.5 (L)	89 (L)	96 (L)
		1,000/1,000	0.12	0.01	0.03	0.10	52.5 (S)	10.5 (S)	28.5 (S)	47 (S)
							95.5 (M)	68.5 (M)	87 (M)	94.5 (M)
							100 (L)	99 (L)	100 (L)	100 (L)
LR	20	500/500	0.09	0.01	0.01	0.01	38 (S)	7 (S)	7 (S)	12 (S)
							88 (M)	62 (M)	63 (M)	68 (M)
							99 (L)	97 (L)	97 (L)	97 (L)
		1,000/1,000	0.14	0.02	0.02	0.05	71 (S)	31 (S)	32 (S)	50 (S)
							99 (M)	96 (M)	96 (M)	98 (M)
							100 (L)	100 (L)	100 (L)	100 (L)
	40	500/500	0.10	0.00	0.00	0.02	46.5 (S)	8.5 (S)	9 (S)	17.5 (S)
							90 (M)	49 (M)	49.5 (M)	67.5 (M)
							100 (L)	90.5 (L)	91.5 (L)	97 (L)
		1,000/1,000	0.14	0.01	0.01	0.04	71.5 (S)	20.5 (S)	22 (S)	43 (S)
							100 (M)	92.5 (M)	93 (M)	98.5 (M)
							100 (L)	100 (L)	100 (L)	100 (L)

(continued)

Table 2. (continued)

	Condition		Type I Error Rate				Power Rate (With DIF Magnitude)			
Method	Test Length	Sample Size	Unadjusted	Bonferroni		BH	Unadjusted (in %)	Bonferroni (in %)	Holms (in %)	BH (in %)
DFIT	20	500/500	0.03	0.01	0.01	0.01	48 (S)	4 (S)	4 (S)	14 (S)
							93 (M)	51 (M)	51 (M)	62 (M)
							97 (L)	76 (L)	76 (L)	80 (L)
		1,000/1,000	0.05	0.01	0.01	0.02	78 (S)	28 (S)	28 (S)	46 (S)
							98 (M)	91 (M)	91 (M)	95 (M)
							98 (L)	96 (L)	96 (L)	96 (L)
	40	500/500	0.03	0.00	0.00	0.01	49 (S)	6 (S)	6 (S)	30 (S)
							94 (M)	52.5 (M)	52.5 (M)	75.5 (M)
							100 (L)	90 (L)	90 (L)	96 (L)
		1,000/1,000	0.04	0.00	0.00	0.01	80.5 (S)	26 (S)	26 (S)	67 (S)
							99.5 (M)	92.5 (M)	92.5 (M)	98.5 (M)
							100 (L)	98.5 (L)	98.5 (L)	99.5 (L)
Lord	20	500/500	0.02	0.00	0.00	0.00	16 (S)	2 (S)	2 (S)	2 (S)
							56 (M)	12 (M)	12 (M)	15 (M)
							84 (L)	48 (L)	48 (L)	50 (L)
		1,000/1,000	0.02	0.00	0.00	0.00	38 (S)	5 (S)	5 (S)	10 (S)
							85 (M)	54 (M)	55 (M)	58 (M)
							98 (L)	90 (L)	90 (L)	92 (L)
	40	500/500	0.02	0.00	0.00	0.00	18 (S)	2 (S)	2 (S)	4 (S)
							18 (M)	2 (M)	2 (M)	4 (M)
							90.5 (L)	55 (L)	54.5 (L)	61.5 (L)
		1,000/1,000	0.03	0.00	0.00	0.00	41 (S)	6.5 (S)	8 (S)	19 (S)
							87.5 (M)	47 (M)	47 (M)	64.5 (M)
							100 (L)	94 (L)	94 (L)	97.5 (L)

Note. DIF = differential item functioning; BH = Benjamini–Hochberg; MH = Mantel–Haenszel method; DFIT = differential functioning item and test; LR = logistic regression; DIF magnitude: Small (S; 0.3), Medium (M; 0.5), and Large (L; 0.7).

Discussion

Conclusion and Significance

In this study, Holm's procedure and the BH false discovery rate were effective in controlling the Type I error rates of MH and LR. In particular, the BH false discovery rate seemed to be the most balanced method in lowering the Type I error rate, compared with the Bonferroni correction and Holm's procedure as previous research also suggested (Penfield, 2001; Thissen et al., 2002).

One negative effect of using such procedures, however, was the decreased power. The analysis of power showed a very strong and consistent pattern that DIF items were detected with (almost) perfect accuracy with the large DIF magnitude. The power was consistently high even after adjustments with this large DIF magnitude. This is important, as in a practical testing situation, large DIF is of primary concern. However, when the DIF magnitude was medium, the power was reduced substantially after three adjustments were applied in most conditions. Even though all three procedures reduced the power when the DIF magnitude was medium, the BH false discovery rate seemed to lose power the least.

Another interesting finding of this study is that the Type I error rate of the DFIT method and of the Lord's chi-square test were well-controlled even before adjustment. Therefore, for the IRT-based tests investigated here, adjustment may not be necessary. Furthermore, it is encouraging that, for DFIT, if the adjustment was applied regardless, the loss of power was not a concern for medium–large DIF, given sufficient sample size (1,000 in each group).

This study has some limitations. First, this study investigated only one aspect of DIF detection practice, that is hypothesis testing. In practice, DIF would be evaluated by both hypothesis testing and effect size. It was our intention in this study, however, to improve the hypothesis testing side of DIF analysis. Second, this study aimed at investigating uniform type of DIF that one group has a consistently better chance of correctly answering an item across the ability range. Further study will also incorporate an additional analysis with nonuniform type of DIF, in which one group does not have a consistently better chance of correctly answering an item.

Future study can evaluate the effectiveness of adjustment procedures along with the use of effect size. MH and LR offer well-established effect size measures (Dorans & Holland, 1993). It should be noted, however, that effect sizes for Lord's chi-square test and DIFT are not currently available. Second, generalization of this study is limited to the conditions investigated here, including the type of DIF (defining DIF magnitude by using *b*-parameter only), test length, sample size, ability distributions, and DIF methods. In particular, further study is needed that includes the conditions where ability distributions differ between the focal and reference groups (i.e., impact) and other popular DIF indices, such as the likelihood ratio test. Despite its limitations, this study demonstrated that the effectiveness of adjustment procedures depended on DIF methods, and, in particular, adjustment might not be necessary for some IRT-based DIF methods.

Appendix

Application of Three Adjustment Procedures (i.e., 15 items, family-wise alpha of .05)

15 items have own associated *p* values. And those 15 *p* values are ordered from the smallest to the largest.

Item	*p value*
1	.001
2	.001
3	.0017
4	.0022
5	.0032
6	.0045
7	.0055
8	.0062
9	.0730
10	.0896
11	.1342
12	.2689
13	.4625
14	.5813
15	.6437

15 items and associated *p* values are ordered from the largest to the smallest. Then *p* values are corrected by *p* value*(total number of items/item's own rank). For example, 8th item, corrected *p* value is .0062*(15/8)=.0116

Family-wise α was corrected by α/(total number of items-item's rank order+1). For example, 8th item, corrected α is .05/(15-8+1)=.0062

Family-wise α was corrected by α/ total number of items .05/15)

Bonferroni Correction

Item	*p-value*	Corrected alpha	
1	.001	.0033	Sig.
2	.001	.0033	Sig.
3	.0017	.0033	Sig.
4	.0022	.0033	Sig.
5	.0032	.0033	Sig.
6	.0045	.0033	
7	.0055	.0033	
8	.0062	.0033	
9	.0730	.0033	
10	.0896	.0033	
11	.1342	.0033	
12	.2689	.0033	
13	.4625	.0033	
14	.5813	.0033	
15	.6437	.0033	

Holm's Procedure

Item	*p-value*	Corrected alpha	
1	.001	.0033	Sig.
2	.001	.0036	Sig.
3	.0017	.0038	Sig.
4	.0022	.0042	Sig.
5	.0032	.0045	Sig.
6	.0045	.0050	Sig.
7	.0055	.0056	Sig.
8	.0062	.0063	Sig.
9	.0730	.0071	
10	.0896	.0083	
11	.1342	.0100	
12	.2689	.0125	
13	.4625	.0167	
14	.5813	.0250	
15	.6437	.0500	

BH False Discovery Rate

Item	*p-value*	Corrected *p* value	
15	.6437	.6437	
14	.5813	.6228	
13	.4625	.5337	
12	.2689	.3361	
11	.1342	.1830	
10	.0896	.1344	
9	.0730	.1217	
8	.0062	.0116	Sig.
7	.0055	.0118	Sig.
6	.0045	.0113	Sig.
5	.0032	.0096	Sig.
4	.0022	.0083	Sig.
3	.0017	.0085	Sig.
2	.001	.0075	Sig.
1	.001	.0150	Sig.

Declaration of Conflicting

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289-300.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probability [Statistical class theory and calculation of probability]. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psycoholgy, 31*, 144-152.

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256.

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

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*, 940-958.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Measurement Methods for the Social Sciences Series). Newbury Park, CA: Sage.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43*, 417-423.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel–Haenszel procedure. In H. Waiter & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Jodoin, M. G., & Gierl, M. L. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel–Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's Differential Functioning of Items and Tests (DFIT). *Educational Measurement*: *Issues and Practice, 27*, 43-50.

Oshima, T. C., Raju, N., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education, 14*, 235-259.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N., Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427-469.

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

Roussos, L. A., & Stout, W. (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*, 215-230.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*(5), 61-84.

Soulakova, J. N. (2009). On identifying effective and superior drug combinations via Holm's procedure based on the Min tests. *Journal of Biopharmaceutical Statistics, 19*, 280-291.

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332-342.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in mutiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77-83.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state to state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42-69.

Zumbo, B. D. (1999). *A handbook on theory and methods of differential item functioning (DIF)* (No. K1A 0K2; pp. 1-57). Ottawa, Ontario, Canada: Human Resources Research and Evaluation, National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.