Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

Gender Differential Item Functioning in Polytomous Items: A Comparison of Three Methods

Consuelo T. Chua
Jose Q. Pedrajita
Kevin Carl P. Santos
University of the Philippines – Diliman

Abstract

The present study compared the consistency of the results of three non-parametric differential item functioning (DIF) techniques – the Cumulative Common Log-Odds Ratio (CCLOR), Standardized Mean Difference (SMD), and the Mantel Test (Mantel) in detecting gender DIF in the *Emotional Quotient Scale* – College Version. The sample comprised 1,229 college students (male = 657; women= 572) from a state university in the Philippines. The agreement of the DIF methods was determined using classification consistency and matching percentages. Results show that CCLOR and Mantel agreed perfectly in detecting gender DIF items. SMD, on the other hand, had a moderate to high agreement with the two other DIF techniques. The agreement among the DIF methods was lower when DIF effect size was considered.

Keywords: DIF, Classification Consistency, Standardized Mean Difference, Common Log-odds Ratio, Mantel Test

Introduction

The integrity of a test depends largely on the quality of its items. A test is valid when it contains items that measure relevant characteristics. On the contrary, when test scores depend on extraneous factors such as group membership (e. g., gender, social status), test bias is said to be present. Bias refers to the presence of systematic error that distorts the outcomes of a test for a

particular group (Camilli & Shepard, 1994; Osterlind, 1983). Bias is a technical term that simply refers to "the consistent distortion of a statistic" (Osterlind, 1983, p. 10) and does not necessarily suggest test unfairness (Penfield & Camilli, 2007).

Statistical methods to detect potentially biased test items were first developed in the 1970s. However, it was only in the 1980s that a 'general statistical framework,' now termed as differential item functioning (DIF), was developed as basis for the analysis of 'item statistical bias' (Penfield & Camilli, 2007, p. 126). The term 'potentially biased' is used to denote that DIF items are not automatically considered biased until the source of differential item functioning is explained.

DIF pertains to the difference in the performance of two or more matched groups (e.g., gender groups, age groups) on a test item. DIF occurs when members of groups who are similar in ability have different chances of obtaining a correct response or score on an item, leading to an unfair advantage for one group over the others (Penfield & Camilli, 2007; Roussos & Stout, 2004). For instance, if women are more likely to obtain a higher score on an essay item compared to men with similar proficiency then gender-based DIF is present. Although several groups may be involved in DIF analysis, two groups are normally compared - the reference group and the focal group. The reference group is the group that the test expects to favor while the focal group is the group that is likely to be disadvantaged by the test.

For dichotomous items, DIF pertains to the difference in the probability of a correct response between two groups with similar proficiency (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). On the other hand, DIF is present in polytomous items when the probability of obtaining an item score differs between two matched groups. At present, there are numerous methods that are available for analyzing DIF for dichotomous items including item response theory (IRT) approaches, proportion-difference approaches, and common-odds ratio approaches (see Penfield & Camilli, 2007). Studies about DIF for dichotomous items are also abundant (e.g., Wiberg, 2009; Pedrajita, 2007; Gibson & Harvey, 2003). However, many tests contain items that are polytomously scored. Compared with dichotomous DIF, polytomous DIF deals with several score levels which makes DIF analysis more complex (Penfield & Camilli, 2007; Potenza & Dorans, 1995).

A number of methods have been proposed for polytomous DIF; several of which are extensions of procedures for dichotomous DIF including mean-difference approaches, multivariate hypergeometric distribution approaches,

and common odds ratio approaches (see Penfield & Camilli, 2007). However the agreement among several techniques for polytomous DIF has not been explored. The standardized mean difference (SMD) for instance is a non-parametric DIF technique that is used by the National Assessment of Educational Progress (NAEP) as an effect size estimator. On the other hand, the cumulative common log-odds ratio (CCLOR) is another polytomous DIF method that is also widely used to determine DIF effect size. The agreement between these two methods has yet to be examined.

Comparing the effect sizes of polytomous DIF techniques is important because DIF outcomes should be interpreted not only in terms of statistical significance but also based on the severity or size of potential bias. This is true because significant DIF outcomes may be observed "even for negligible departures from the null hypothesis" and small DIF effect sizes do not normally possess "practical value" (Meyer, Huynh, & Seaman, 2004, p. 335). Another DIF technique that is widely used is the Mantel Test (Mantel, 1963). This method is an extension of the Mantel-Haenszel procedure that has been widely used in previous studies. Thus, it would be beneficial to compare the Mantel test with SMD and CCLOR. In addition, these three methods are non-parametric statistics that are easier to understand compared to more complex DIF techniques such as IRT-based methods. Furthermore, parametric DIF techniques require that certain conditions are met (e.g., sample size) which limit their use in practical settings (Penfield, Glacobbi, & Myers, 2007). In contrast, the Mantel test, SMD, and CCLOR are non-parametric tests that do not require stringent assumptions. For this reason, these methods can be easily employed in many actual test situations. However, the performance of these methods in detecting DIF items has not been investigated empirically in the literature. This study hopes to fill in this gap by comparing the ability of these three methods in identifying DIF items.

The Mantel, SMD, and CCLOR tests have been applied in several studies across various tests to measure different types of DIF. For instance, CCLOR was applied by Penfield, Giacobbi, and Myers (2007) to detect gender DIF in the Exercise Imagery Inventory using total score of the relevant subscales as matching variable. Two out of the 19 items in the scale, one with moderate DIF and another with large DIF, were flagged as functioning differently between genders. In another study, standardized CCLOR was used as DIF effect size estimator for Mantel-Haenzel, alongside Ordinal Logistic Regression to identify cross-cultural DIF in the Student Questionnaire of the Program for International Student. The results yielded 14 items that were

similarly flagged as having either medium or large DIF using both procedures (Padilla, Baena, Hidalgo, & Sireci, 2011). On the other hand, the study of Wetzel and Hell (n.d.) examined gender DIF on the Allgemeiner Interessen-Struktur-Test [General Interest Structure Test] using CCLOR and IRT. DIF analyses show that the two procedures generally agree in flagging DIF items. An examination of the foregoing studies shows that CCLOR can be effectively used to identify DIF in scaled items, especially when DIF effect size is sought. CCLOR also has a good agreement with other DIF methods.

Similar with CCLOR, SMD has also been used to measure DIF in a variety of tests. Fletcher (2008) used SMD to measure the DIF effect size of the Likelihood Ratio Test in identifying ethnicity–based DIF on the Risk Perception Survey for Mellitus (RPS-DM). The outcomes of the analysis showed that the RPS-DM contained five items with strong DIF and one item with weak DIF. Similarly, Schwarz, Rich, and Podrabsky (2003) applied SMD together with the Linn-Harnisch procedure to examine DIF based on mode of test administration (on-line or paper and pencil). Two instruments were used for the analysis including an aptitude test and knowledge test on Reading, Mathematics and Language. A few items were flagged as having small to moderate DIF, some against the on-line group and others against the paper and pencil group. In another study, SMD was also used together with the Mantel test to identify gender DIF on an ordinal self-concept scale. Results showed that 42% of the items contained DIF which vary in both direction and magnitude (Young & Sudweeks, 2005). The aforementioned studies demonstrated that SMD can be used to measure DIF across different test types (e.g., cognitive and affective) and among varied DIF comparisons (e.g., gender, ethnicity, and test administration).

Aside from the previously mentioned study of Young and Sudweeks (2005), the Mantel procedure has also been applied to detect DIF in other contexts. For example, Henderson (2001) used Mantel together the Mantel-Haenszel (MH), SIBTEST, and Poly-SIBTEST to determine gender DIF on an academic high school exit test. The matching variable involved the corresponding total test score for each item. DIF results revealed that about 15% of the dichotomous items had DIF, while more DIF items were detected among polytomous items. The Mantel also displayed good agreement with MH in detecting DIF items. In another study, Cameron, Scott, Adler, and Reid (2014) identified age and gender-DIF on the Hospital Anxiety Depression Scale using the Mantel, ordinal logistic regression, and Rasch Analysis. The three methods similarly flagged three age-related DIF items, showing the general

consistency among the methods in detecting DIF. An inspection of the abovementioned studies illustrates that the Mantel is sensitive in detecting various forms of DIF and has high consistency with other DIF techniques.

Considering the aforementioned studies on CCLOR, SMD, and Mantel, it is worthwhile to compare the results of the three tests because SMD and CCLOR are both established effect size estimators that may be effectively used in combination with Mantel. Conducting this comparison would provide both statistical and practical interpretations of DIF, which allows for a more accurate interpretation of DIF results. In addition, the three methods have similar characteristics, including their applicability to ordinal data and the use of sum of scores as matching variable that facilitates DIF comparison on a same given test. Therefore, the purpose of this study was to compare the agreement or consistency among three polytomous DIF statistics – the SMD, MANTEL, and CCLOR in detecting gender DIF items in a polytomous emotional quotient test. The study specifically aimed to: (1) examine the consistency among the methods in detecting gender DIF based on statistical significance; and (2) determine the consistency between CCLOR and SMD in classifying DIF items based on substantial significance and effect size. Comparison based on substantial DIF and effect size was only possible for CCLOR and SMD because MANTEL by itself does not produce a measure of DIF effect size.

Method

Participants

The sample comprised 1,229 undergraduate students (657 are males; 572 are females) from the University of the Philippines - Diliman. The mean age of the respondents was 20. The respondents were mostly third to fifth year college students coming from varied science-related courses. Only respondents who were at least in third year college were selected because there were items in the scale that may not be applicable to younger students.

Research Instrument

The DIF analysis was performed using the *Emotional Quotient Scale* – college version *(EQS-C)* which was developed by Marquez (2002) and distributed by MAVEC Specialist Foundation Inc. An emotional quotient scale was chosen because studies have consistently shown differences between

genders on the construct 'emotional intelligence' (e.g., Tapia & Marsh, 2006; Rooy, Alonso, & Viswesvaran, 2005). It would therefore be significant to determine if gender DIF has a contribution to such measurement differences.

The *EQS-C* is a 140-item emotional intelligence test for college students and a modification of the *EQS* test for adults and employees (*EQS-AE)*. The scale has five response options ranging from *very true of me* to *very untrue of me*. The test has ten subscales: adaptability, communication (15 items), confidence (20 items), decision-making (15 items), empathy (15 items), interpersonal skills (23 items), motivation (7 items), innovation (6 items), teamwork (21 items), and trustworthiness (3 items). However, three subscales - innovation, motivation, and trustworthiness were excluded from the analysis because their Cronbach's alpha estimates were lower than r=.70. All the other subscales have moderate to high Cronbach's alpha (r=.75 to r=.86).

Data Collection Procedures

After obtaining the necessary permits from the concerned officials and faculty members of the University, the *EQS-C* was administered to 55 separate classes of students. The following procedures were followed for each data collection session. First, the students were requested to participate in the study by answering the *EQS-C*. They were also informed that participation was voluntary, and those who agreed to participate were requested to accomplish consent forms. Next, the students were provided with the test materials and were given directions on answering the test. Finally, the students answered the *EQS-C* and submitted their answer sheets. Each data collection session lasted for 30 minutes on the average. Data were collected during the first quarter of 2013.

Data Analysis

The initial step of data analysis involved detecting the gender DIF items in the *EQS-C* using CCLOR, MANTEL, and SMD. This procedure served as the basis for determining the agreement among the three DIF methods. Gender DIF detection for all three methods was implemented using a procedure similar to the one outlined by Penfield and Camilli (2007). The first step involved establishing the sum of scores for each subscale as the matching variable. The next step involved establishing the reliability of each matching variable by

computing for the Cronbach's alpha coefficient for each subscale. The third procedure concerned conducting a t-test to ensure that the scores of the two gender groups on the different subscales were matched. The fourth step concerned stratifying the respondents into three equally-spaced ability levels based on their total score per subscale. The final step involved applying three statistical techniques - CCLOR (Penfield & Algina, 2003), SMD, and MANTEL (Mantel, 1963) to detect gender DIF on each subscale or dimension of the test. Separate DIF analyses were conducted for each subscale of the *EQS-C* because DIF assessment assumes that all items "measure the same dimension." Therefore, in multidimensional tests such as the *EQS-C*, it is essential to decompose scores "into more homogenous subscores" to ensure the validity of the DIF outcomes (Dorans & Holland, 1993 as cited in Potenza & Dorans, 1995, p. 32). An item is flagged as DIF based on statistical significance if the DIF tests were significant at $\alpha = .05$. On the other hand, an item was flagged as having *substantial DIF* when the DIF effect size is at least moderate or large.

Cumulative Common Log-odds Ratio(CCLOR). The cumulative common log-odds ratio determines the difference between the focal and reference groups in terms of the odds of exceeding each category of the studied item, while controlling for target trait (Penfield, Giacobbi, & Myers, 2007). The cumulative common odds ratio was proposed by Liu & Agresti (1996) and applied to the detection of DIF of polytomous items by Penfield & Algina (2003).

CCLOR is computed based on the average estimated odds ratio of all the response categories of the item within each ability level. The average of the odds ratio is called the *cumulative common odds ratio estimator*. This estimator cannot have negative values and as such, its natural logarithm called the *cumulative common log-odds estimator* is taken to allow both negative and positive values (Penfield, Giacobbi, & Myers, 2007; Penfield & Algina, 2003).

The null hypothesis is that the odds of exceeding each response level of an item is the same for the two groups (Penfield, Giacobbi, & Myers, 2007). A zero *CCLOR* value indicates the absence of DIF; while values of *CCLOR* < 0 shows DIF in favor of focal group (males) and values of *CCLOR* > 0 shows DIF in favor of reference group (females). The DIF effect size of *CCLOR* was determined based on a classification scheme proposed by Penfield (2007), as shown below.

- Category AA (small) – when either *CCLOR* is not significantly different from zero or $|CCLOR| < .43$

- Category BB (moderate) – when *CCLOR* is significantly different from zero and $|CCLOR| \geq .43$ and either $|CCLOR| < .64$ or $|CCLOR|$ is not significantly greater than .43
- Category CC (large) - when $|CCLOR|$ is significantly greater than 0.43 and $|CCLOR| \geq .64$

The DIFAS 5.0 program, developed by Randall Penfield was used to compute for the cumulative common log-odds ratio, *CCLOR*; the standard error of *CCLOR*, and the standardized *CCLOR* (Z statistic). Corresponding p-values were computed using Excel in order to test the null hypothesis that *CCLOR* =0. A modified standardized *CCLOR* was also computed for each item in order to test whether the value of *CCLOR* is not significantly greater than .43. The formula used for computing the modified test statistic was [|CCLOR|-.43]/standard error.

The Mantel Test (MANTEL). The Mantel Test (Mantel, 1963), an extension of the Mantel-Haenszel technique is used to test the association between two matched groups on ordinal items (Welch & Hoover, 1993). It tests the null hypothesis of no association between group and the response variable.

An item is flagged as having DIF under the Mantel Test if the null hypothesis is rejected at α = .05. MANTEL by itself does not provide information on the direction of DIF and a classification scheme to categorize the effect size of DIF. The GMHDIF program developed by Fidalgo (2011) was used to detect DIF using this procedure.

Standardized Mean Difference (SMD). The standardized mean difference is a descriptive index that compares the item means of the focal and reference groups, after adjusting for differences in the distribution of members of the groups across the values of the matching variable (Zwick, Thayer & Mazzeo, 1997). The null hypothesis is that the population value of the standardized mean difference is zero. An SMD index of zero pertains to the absence of DIF. In general, positive SMD values correspond to DIF in favor of the reference group while negative SMD values correspond to DIF in favor of the focal group (Penfield, Giacobbi, & Myers, 2007). However, for this study, the SMD program was written such that positive SMD values pertain to DIF in favor of males (the focal group), whereas negative SMD values pertain to DIF in favor of females (the reference group). The R software was utilized to run the test using the SMD R Script provided in the study of Wood (2011).

The National Assessment of Educational Progress (NAEP) categorization scheme for classifying the effect size of DIF was used to interpret the size of DIF for standardized mean difference (John Donoghue, Personal Communication). The classification is shown below.

- Category AA (small or negligible) - if Mantel Chi-square p-value < 0.05 and |SMD/SD| ≤ 0.17
- Category BB (moderate) - if Mantel Chi-square is significant at p-value < 0.05 and |SMD/SD| > 0.17
- Category CC (large) – if Mantel Chi-square is significant at p-value < 0.05 and |SMD/SD| > 0.25

SMD refers to the standardized mean difference index while SD pertains to the group standard deviation of the item score. For the purposes of this study, only the value of SMD and not of Mantel was considered in determining the effect size of SMD. Items with significant SMD values at α =.05 were considered as DIF items and the value of |SMD/SD| as previously mentioned was used to classify DIF items.

Consistency Among Polytomous DIF Methods. The agreement among CCLOR, MANTEL, and SMD was determined using classification consistency and matching percentage. Classification consistency involves the simple procedure of comparing the number of DIF items consistently detected by the methods. On the other hand, matching percentages were computed by obtaining the ratio between the number of items detected using both procedures under comparison and the number of items detected using at least one procedure (Wiberg, 2009). The resulting matching percentages were interpreted as follows.

- High Matching Percentage - 75 to100%
- Moderate Matching Percentage - 50% to 74%
- Low Matching Percentage - less than 50 %

The consistency among the DIF methods was determined based on three DIF interpretations – DIF based on statistical significance, DIF based on substantial significance, and DIF based on effect size. The consistency of the three methods in detecting DIF items based on statistical significance was determined by considering all flagged DIF items regardless of effect size. On the other hand, the agreement between CCLOR and SMD to detect DIF based

on substantial significance was determined by considering only DIF items that have at least a moderate effect size or classified as category BB or CC. Finally, the consistency of the methods in flagging DIF based on effect size was determined by obtaining the frequency DIF items that were consistently classified as small, moderate, or large by the two procedures. Only CCLOR and SMD were considered in the DIF comparison based on substantial DIF and effect size because as previously mentioned, MANTEL by itself does not produce a measure of DIF effect size.

Results and Discussion

Gender DIF Items in the Emotional Quotient Scale

Cumulative Common Log-Odds Ratio (CCLOR). CCLOR detected 47 gender DIF items based on statistical significance in the *EQS-C*. Table 1 shows the gender DIF items that were detected for each subscale of the test and the corresponding DIF effect sizes. Twenty-four of these items were potentially biased toward males, which meant that males have a greater probability of attaining a higher score in these items compared to females with similar EQ levels. On the other hand, 23 DIF items were in favor of females. For these items, females have a higher chance of receiving a higher score on the items compared to males who have the same EQ level. However, out of the 47 gender DIF items, only 12 had substantial (moderate or large) amounts of DIF (Table 2). Four of these items had large DIF effect sizes, while eight had moderate DIF effect sizes. The rest of the DIF items only had small or negligible DIF which is not sufficient to conclude the presence of DIF.

Standardized Mean Difference (SMD). SMD flagged 45 gender DIF items in the seven subscales of the *EQS-C* (Table 1). Most of these items, 25 in all, were potentially biased towards males. For these items, males had a possible unfair advantage of obtaining higher scores compared to females with similar abilities on the corresponding subscale to which the items belong. On the other hand, 20 items were biased towards females which meant that females had a higher probability of obtaining higher scores in these items compared to males. Among the 45 DIF items detected, only 15 had substantial amounts of DIF (Table 2). Nine items had large DIF, while seven had moderate DIF. The rest of the items had small or negligible DIF.

The Mantel Test (MANTEL). The Mantel Test detected 47 gender DIF items (Table1). MANTEL does not provide a value to determine the direction of DIF – whether the item is in favor of the reference group or the focus group . Furthermore, MANTEL by itself does not provide an effect size estimate for the severity of DIF. Thus, when employing MANTEL for DIF detection, other DIF techniques that can serve as effect size estimators. Both CCLOR and SMD can serve this purpose.

All in all, the three DIF procedures detected 50 gender DIF items or 40% of the 124 items in the seven subscales of the Emotional Quotient Scale. However, only 16 out of the 50 items had substantial DIF or at least moderate DIF effect size. Further, of the 50 DIF items, 42 items were detected by all three methods. A greater number of DIF items (27 items) were found to be in favor of males compared to only 23 DIF items that were favorable to females. This is an unexpected result given that EQ scales mostly produce scores that are favorable to females (e.g., Rooy, Alonso, & Viswesvaran, 2005; Day & Carol, 2004).

The CCLOR and MANTEL procedures were equally sensitive in detecting DIF based on statistical significance and flagged 47 items each. The standardized mean difference procedure was slightly more conservative and detected only 45 DIF items. However, although CCLOR was more sensitive in flagging DIF items than SMD based on statistical significance, the former was a more conservative effect size estimator. For instance, items 6, 10, 18, and 98 all had moderate DIF under the SMD procedure but only had small DIF under the CCLOR procedure.

Table 1

Gender DIF Items in the Emotional Quotient Scale based on Statistical Significance

	CCLOR			SMD			MANTEL	
Item	CCLOR	p-value	In Favor of	SMD	p-value F	In Favor of	MANTEL	p-value
				Adaptability				
7	0.296	0.008	F	-0.125	0.008	F	7.027	0.008
18	0.376	0.001	F	-0.134	0.001	F	10.571	0.001
61	-0.326	0.003	M	0.132	0.003	M	8.688	0.003
90	-0.56	0.000	M	0.287	0.000	M	27.439	0.000
93	-0.275	0.016	M	0.099	0.014	M	5.871	0.015
125	0.316	0.010	F	-0.093	0.011	F	6.633	0.010
127	0.503	0.000	F	-0.187	0.000	F	18.810	0.000
128	0.346	0.002	F	-0.161	0.002	F	10.024	0.002
				Communication				
38	0.345	0.004	F	-0.140	0.004	F	8.339	0.004
47	0.439	0.003	F	-0.056	0.040	F	8.776	0.003
60	0.351	0.004	F	-0.119	0.005	F	8.080	0.005
78	-0.231	0.041	M	0.107	0.042	M	4.152	0.042
82	0.338	0.008	F	-0.097	0.007	F	7.168	0.007
84	NS	NS	NS	0.094	0.047	M	NS	NS
89	-0.305	0.007	M	0.134	0.009	M	7.159	0.008
94	0.252	0.040	F	-0.099	0.026	F	4.255	0.039
98	-0.402	0.000	M	0.174	0.000	M	12.109	0.001
				Confidence				
6	0.518	0.000	F	-0.161	0.000	F	18.502	0.000
9	-0.804	0.000	M	0.352	0.000	M	51.637	0.000
48	-0.332	0.003	M	0.131	0.003	M	8.660	0.003
49	0.721	0.000	F	-0.237	0.000	F	34.266	0.000
50	-0.238	0.028	M	0.112	0.028	M	4.861	0.028
95	0.586	0.000	F	-0.203	0.000	F	25.365	0.000
99	0.399	0.001	F	-0.133	0.001	F	11.772	0.001
113	-0.464	0.000	M	0.191	0.000	M	17.884	0.000
131	-0.36	0.002	M	0.133	0.002	M	9.666	0.002
134	0.326	0.004	F	-0.126	0.004	F	8.320	0.004
				Decision-Making				
73	-0.251	0.024	M	0.097	0.025	M	5.088	0.024
120	-0.424	0.000	M	0.188	0.000	M	14.741	0.000
123	0.356	0.004	F	-0.123	0.002	F	8.711	0.003
136	NS	NS	NS	0.081	0.049	M	NS	NS
				Empathy				
56	NS	NS	NS	0.124	0.046	M	NS	NS
1	0.294	0.013	F	-0.125	0.009	F	6.266	0.012
				Interpersonal Skills				
13	0.594	0.000	F	-0.341	0.000	F	27.502	0.000
32	-0.242	0.042	M	NS	NS	NS	4.127	0.042
54	-0.29	0.012	M	0.121	0.013	M	6.398	0.011
70	-0.589	0.000	M	0.215	0.000	M	19.945	0.000
76	0.65	0.000	F	-0.425	0.000	F	30.747	0.000
79	0.304	0.043	F	NS	NS	NS	4.111	0.043
114	-0.262	0.035	M	0.081	0.045	M	4.504	0.034
116	-0.3	0.020	M	0.087	0.016	M	5.381	0.020
135	-0.291	0.014	M	0.118	0.019	M	6.014	0.014
				Teamwork				
10	-0.4	0.000	M	0.178	0.000	M	12.675	0.000
31	-0.767	0.000	M	0.348	0.000	M	47.398	0.000
57	0.366	0.004	F	-0.094	0.001	F	8.202	0.004
58	-0.269	0.011	M	0.148	0.013	M	6.407	0.011
67	0.242	0.032	F	NS	NS	NS	4.516	0.034
104	0.305	0.015	F	NS	NS	NS	6.013	0.014
118	-0.351	0.004	M	NS	NS	NS	8.418	0.004
132	-0.283	0.027	M	0.072	0.025	M	4.889	0.027

Note: F=females; M=males; NS = Not Significant

Table 2
Items that have Substantial DIF using CCLOR and SMD

Item	CCLOR				SMD			
	CCLOR	*CCLOR* Modified	p-value (*CCLOR* ≥0.43)	Size of DIF	SMD	SD	SMD/ SD	Size of DIF
			Adaptability					
18	0.376	-0.466	0.679	Small	-0.134	0.774	-0.174	Mod
90	-0.560	1.215	0.112	Mod	0.287	1.032	0.278	Large
127	0.503	0.624	0.266	Mod	-0.187	0.728	-0.257	Large
			Communication					
47	0.439	0.060	0.476	Mod	-0.056	0.546	-0.103	Small
98	-0.402	-0.243	0.596	Small	0.174	0.954	0.183	Mod
			Confidence					
6	0.518	0.727	0.234	Small	-0.161	0.717	-0.224	Mod
9	-0.804	3.369	0.000	Large	0.352	0.954	0.369	Large
49	0.721	2.347	0.009	Large	-0.237	0.766	-0.309	Large
95	0.586	1.333	0.091	Mod	-0.203	0.810	-0.251	Large
113	-0.464	0.312	0.378	Mod	0.191	0.918	0.209	Mod
			Decision-making					
120	-0.424	-0.054	0.522	Small	0.188	1.003	0.188	Mod
			Interpersonal Skills					
13	0.594	1.426	0.077	Mod	-0.341	1.130	-0.302	Large
70	-0.589	1.205	0.114	Mod	0.215	0.857	0.251	Large
76	0.650	1.849	0.032	Large	-0.425	1.273	-0.334	Large
			Teamwork					
10	-0.400	-0.270	0.607	Small	0.178	0.933	0.190	Mod
31	-0.767	2.982	0.001	Large	0.348	0.927	0.375	Large

Note: DIF is substantial if test statistic is significant and DIF effect size is at least moderate; Mod=moderate; CCLOR Modified is the test statistic used to determine whether CCLOR is significantly greater than or equal to .43; SD = standard deviation

Detection Consistency among DIF Methods Based on Statistical Significance

The number of DIF items consistently detected by CCLOR, SMD, and MANTEL and the matching percentages of agreement among the methods based on statistical significance are shown in Table 3. MANTEL and CCLOR had a perfect agreement (100% matching percentage) in all subscales. The two methods commonly flagged a total of 47 items across the subscales, including items 7, 18, 61, 90, 93, 125, 127, 128, 38, 47, 60, 78, 82, 89, 94, 98, 6, 9, 48, 49,

50, 95, 99, 113, 131, 134, 73, 120, 123, 1, 13, 32, 54, 70, 76, 79, 114, 116, 135, 10, 31, 57, 58, 67, 104, 118, and 132 (Table 1). On the other hand, the comparison between CCLOR and SMD; and MANTEL and SMD both resulted to an agreement ranging from 50% to 100% which is moderate to high. CCLOR and SMD consistently detected all items except for items 84, 136, 56, 32, 79, 67, 104, and 118.

The same items were consistently detected by MANTEL and SMD. The perfect consistency between MANTEL and CCLOR shows that the two methods are compatible and can be used together to check the validity of DIF outcomes. The outcomes also suggest that caution should be used when interpreting items that are flagged as DIF by only one of the methods. Furthermore, the compatibility of the two methods provides a good indication that CCLOR can effectively serve as an effect size estimator for MANTEL. However, the use of SMD as effect size estimator for MANTEL is also appropriate especially if a more conservative DIF outcome is sought. The two methods can serve as check and balance for DIF test of significance.

In general, the comparisons between methods yielded average to high matching percentages, ranging from 50% to 100%. The agreement among all three methods was also generally high. All three methods consistently detected 42 out of the 50 DIF items in the seven subscales including items 7, 18, 61, 90, 93, 125, 127, 128, 38, 47, 60, 78, 82, 89, 94, 98, 6, 9, 48, 49, 50, 95, 99, 113, 131, 134, 73, 120, 123, 1, 13, 54, 70, 76, 114, 116, 135, 10, 31, 57, 58, and 132.

Table 3

Classification Consistency and Matching Percentage of the Three Methods in Detecting DIF Items based on Statistical Significance

	Adaptability		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	8 (100%)	8	
Mantel	8 (100%)	8 (100%)	8
	Communication		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	8 (89%)	9	
Mantel	8 (100%)	8 (89%)	8
	Confidence		
	CCLOR	SMD	Mantel
CCLOR	10		
SMD	10 (100%)	10 (100%)	
Mantel	10 (100%)	10 (100%)	10
	Decision-Making		
	CCLOR	SMD	Mantel
CCLOR	3		
SMD	3 (75%)	4	
Mantel	3 (100%)	3 (75%)	3
	Empathy		
	CCLOR	SMD	Mantel
CCLOR	1		
SMD	1 (50%)	2	
Mantel	1 (100%)	1 (50%)	1
	Interpersonal Skills		
	CCLOR	SMD	Mantel
CCLOR	9		
SMD	7 (78%)	7	
Mantel	9 (100%)	7 (78%)	9
	Teamwork		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	5 (62.50%)	5	
Mantel	8 (100%)	5 (63%)	8

Note: DIF items include all items that were flagged based on significant p-values

Consistency between CCLOR and SMD in Detecting Items with Substantial DIF

The classification consistency and matching percentages between SMD and CCLOR in detecting items with substantial DIF are presented in Table 4. As shown, the matching percentages between SMD and CCLOR varied according to subscale, which ranged from 0 to 100 %. The two methods consistently detected 10 out of the 16 items with substantial DIF (62.50% agreement) in all subscales combined, including items 90, 127, 9, 49, 95, 113, 13, 70, 76, and 31.

Higher matching percentages were found in subscales with more substantial DIF items such as confidence and adaptability. On the other hand, 0% matching percentages were found in subscales such as communication and decision-making, wherein only one item with substantial DIF was detected. Logically, if the two DIF procedures did not agree on the single substantial DIF item, then the matching percentage would automatically be zero.

The consistency between CCLOR and SMD was lower when substantial DIF was considered compared to when DIF was detected based only on tests of significance. This outcome is expected given that another criterion, DIF effect size, is also considered in the comparison. This means that even if the methods agree in terms of flagging DIF based on statistical significance, the severity of DIF detected does not necessarily coincide.

The outcomes show that the agreement among methods is not always perfect especially when substantial DIF is considered. This outcome emphasizes even more the importance of using more than one method in detecting DIF in order to validate the accuracy of DIF results especially when the size of DIF outcome is small. Items should only be interpreted as DIF when these are flagged by both methods.

Table 4
Classification Consistency and Matching Percentage of SMD and CCLOR in Detecting Items with Substantial DIF

	Adaptability	
	CCLOR	SMD
CCLOR	2	
SMD	2 (66.67%)	4
	Communication	
	CCLOR	SMD
CCLOR	1	
SMD	0 (0%)	1
	Confidence	
	CCLOR	SMD
CCLOR	5	
SMD	5 (100%)	5
	Decision-making	
	CCLOR	SMD
CCLOR	0	
SMD	0 (0%)	1
	Interpersonal Skills	
	CCLOR	SMD
CCLOR	3	
SMD	3 (100%)	3
	Teamwork	
	CCLOR	SMD
CCLOR	1	
SMD	1 (50%)	2

Note: DIF is substantial if test statistic is significant at α=.05 and DIF effect size is moderate or large

Consistency between CCLOR and SMD in Classifying Statistically Significant DIF Items based on Effect Size

The consistency between cumulative common log odds ratio and standardized mean difference in classifying statistically significant DIF items based on the effect size of DIF is shown in Table 5. Across all subscales, the two DIF methods had a higher consistency in classifying small DIF compared to moderate and large DIF effect sizes. The two methods consistently classified 26 out of the 39 items (64.10%) with small DIF using either procedure, including items 7, 61, 93, 125, 128, 38, 60, 78, 82, 89, 94, 48, 50, 99, 131, 134, 73, 123, 1, 54, 114, 116, 135, 57, 58, and 132.

CCLOR and SMD had a moderate level of consistency in classifying items with large DIF. The two methods consistently classified 4 out of the 8 items with large DIF (50% agreement), including items 9, 49, 76, and 31. The lowest agreement concerned classification for moderate DIF items. Here, the consistency between CCLOR and SMD was only 18.18% or 2 out of 11 moderate DIF items that were detected by either method. Only items 6 and 113 were consistently classified as moderate by the two methods. Some items that were classified as moderate by SMD were classified only as small by CCLOR. On the other hand, some items that were classified as large by SMD were only classified as moderate by CCLOR. Generally, CCLOR and SMD are not consistent in classifying the size or severity of DIF especially for small and moderate DIF.

Summary, Conclusions, and Recommendations

The present study compared the consistency among DIF methods in detecting potentially biased items in a polytomous scale. The detection consistencies of the methods based on statistical significance, substantial significance, and effect size classification were examined. In order to address the purpose of the study, the *Emotional Quotient Scale* was administered to college students. The outcomes of the test were subjected to DIF analysis using CCLOR, SMD, and MANTEL. The DIF agreement among the three methods was compared using classification consistency and matching percentages.

Table 5

Number of Items Consistently Classified by CCLOR and SMD based on DIF Effect Size

	SMD				
	Adaptability				
CCLOR	Small	Mod	Large	NDIF	Total
Small	4	2			6
Mod			2		2
Large					
NDIF					
Total	5	1	2	0	8
	Communication				
CCLOR	Small	Mod	Large	NDIF	Total
Small	6	1			7
Mod	1				1
Large					
NDIF	1				1
Total	8	1			9
	Confidence				
CCLOR	Small	Mod	Large	NDIF	Total
Small	5	1			6
Mod		1	1		2
Large			2		2
NDIF					
Total	5	2	3		10
	Decision-making				
CCLOR	Small	Mod	Large	NDIF	Total
Small	2	1			3
Mod					
Large					
NDIF	1				1
Total	3	1			4
	Empathy				
CCLOR	Small	Mod	Large	NDIF	Total
Small	1				1
Mod					
Large					
NDIF	1				1
Total	2				2
	Interpersonal Skills				
CCLOR	Small	Mod	Large	NDIF	Total
Small	4				4
Mod			2		2
Large			1		1
NDIF	2				2
Total	6		3		9
	Teamwork				
CCLOR	Small	Mod	Large	NDIF	Total
Small	3	1		3	7
Mod					
Large			1		1
NDIF					
Total	3	1	1	3	8

Note: NDIF = Not DIF

A total of 50 DIF items (40% of the 124 items) were detected by all three methods in the seven subscales of the *EQ Scale* that were included in the study. Of the 50 items, 42 items were flagged by all three methods. However, only 16 items had substantial DIF or an effect size of at least moderate. CCLOR and Mantel were equally sensitive in detecting DIF items and flagged 47 items each while SMD detected only 45 items.

The CCLOR, MANTEL, and SMD had moderate to high levels of consistency in detecting gender DIF. CCLOR and MANTEL had a perfect agreement in detecting gender DIF across all subscales that were investigated. On the other hand, the comparison between SMD and the two other methods yielded a moderate to high levels of consistency. The consistency between CCLOR and SMD in detecting DIF was lower when substantial DIF was considered. Finally, the classification consistency between CCLOR and SMD in classifying DIF based on effect size showed a higher agreement in classifying small DIF items compared to moderate and large DIF. Moderate DIF items were the least consistently classified by the two methods.

The present study provided several contributions for educators and test developers. Primarily, it explained the agreement among three non-parametric polytomous DIF techniques, not only in terms of statistical significance but also in terms of effect size classification. This is important because in reality, most data do not satisfy the conditions of parametric tests. Thus, a look into non-parametric DIF methods such as GMH, CCLOR, and SMD provide wider applications. Furthermore, a comparison of effect size measures between DIF techniques is helpful because interpretations of DIF that combine both statistical and practical significance provide a more accurate interpretation of item bias.

The study provides the following practical recommendations when performing DIF analysis. First, the study established that MANTEL and CCLOR yield very high or even perfect agreement in detecting DIF. Thus, these two procedures can be effectively used together when detecting DIF and classifying the size of DIF. Flagging an item using both procedures provides more confidence on the validity of the outcomes. Further, since user-friendly point-and-click programs such as GMHDIF and DIFAS are available, then the two procedures can be more practical to use compared to SMD which needs to be programmed in R or other software such as SAS. Penfield, Giacobbi, and Myers (2007) gave the same recommendation about the practicality of DIFAS for CCLOR. Further, since CCLOR, MANTEL, and SMD showed relatively high agreement in detecting DIF, caution should be made when interpreting

DIF items that are flagged by only one of these procedures. One suggestion is that an item should be flagged by at least two of the three procedures to be considered as DIF.

The study also provides suggestions for future research. The usual practice in DIF studies is to consider only moderate and large DIF items as biased. However, it has not been established yet whether or not large DIF items, by inspection, have more 'biased' content than items with small or moderate DIF. Are items with large DIF really more biased that small and moderate DIF items based on qualitative examination? It would be helpful to conduct studies that qualitatively compare the bias content of DIF items with different effect sizes. This way, one can determine whether it is justifiable to include only large DIF items when detecting DIF items.

References

Cameron, I. M., Scott, N. W., Adler, M., & Reid, I. C. (2014). A comparison of three methods of assessing differential item functioning (DIF) in the Hospital Anxiety Depression Scale: ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. *Qual Life Res, 23,* 2883-2888. doi: 10.1007/s11136-014-0719-3.

Camilli, G., & Shepard, L. (1994). *Methods for Identifying Biased Items.* Thousand Oaks:Sage Publications.

Day, A., & Carrol, S. (2004). Using an ability-based measure of emotional intelligence to predict individual performance, group performance, and group citizenship behaviors. *Personality and Individual Differences. 36*, 1443-1458.

Fidalgo, A. (2011). GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using Generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement, 35* (3), 247-249.

Fletcher, J. (2008). *Detecting Differential Item Functioning (DIF) in the Diabetes Risk Perception Survey.* (Doctoral Dissertation). Retrieved from https://fordham.bepress.com/dissertations/AAI3353768/.

Gibson, S. G., & Harvey, R. J. (2003). Gender and ethnicity based differential item Functioning on the armed services vocational aptitude battery. *Equality, Diversity, and Inclusion: An International Journal. 22*(4), 1-15.

Henderson, D. L. (2001). *Prevalence of gender DIF in mixed format high school exit examinations.* Retrieved from https://files.eric.ed.gov/fulltext/ED458284.pdf.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Liu, I-M., & Agresti, A. (1996). Mantel-Haenszel-Type inference for cumulative odds ratios with a Stratified ordinal response. *Biometrics, 52*(4), 1223-1234.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association, 58*, 690-700.

Marquez, A. T. (2002). *Emotional Quotient Scale Manual.* Quezon City: Mavec Specialists Foundation Inc.

Meyer, J., Huynh, H., & Seaman, M. (2004). Exact small-sample differential item functioning Methods for polytomous items with illustration based on attitude survey. *Journal of Educational Measurement, 41*, 331-344.

Osterlind, S. (1983). *Test Item Bias.* Beverly Hills: Sage Publications.

Padilla, J. L, Baena, I. B., Hidalgo, M. D., & Sireci, S. G. (October 2011). *Cognitive interviewing evidence on DIF in Polytomous Items of the Student Questionnaire of the PISA.* Paper presented at the 42th Annual Conference of the Northeastern Educational Research Association, Rocky Hill, USA. Retrieved from http://digibug.ugr.es/bitstream/handle/10481/24229/Padilla_Benitez_Hidalgo_Sireci_NERA2011.pdf;jsessionid=620721BEE5F1A5E9A5AF798A529E26CA?sequence=1

Pedrajita, J. O. (2007). *Item Bias Elimination Models for Test Validity and Reliability.* Unpublished doctoral dissertation, University of the Philippines, Diliman.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 335-355.

Penfield, R. D., & Algina, J. (2003). Applyling the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penfield, R. D., & Camilli G. (2007). Differential item functioning and item bias. In C. Rao & S. Sinharay (Eds.), *Handbook of Statistics Psychometrics* (Vol. 26, pp. 125-167). Amsterdam: Elsevier.

Penfield, R. D., Giacobbi, P. R., & Myers, N. D. (2007). Using the cumulative log-odds ratio to identify differential item functioning of rating scale

items in the exercise and sports sciences. *Research Quarterly for Exercise and Sport, 78(*5), 451–464.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomous scored items: a framework for classification and evaluation. *Applied Psychological Measurement. 19*(1), 23 – 37.

Rooy, D. L., Alonso, A., & Viswesvaran, C. (2005). Group differences in emotional intelligence scores: theoretical and practical implications. *Personality and Individual Differences.38*, 689-700.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF items and testing DIF hypotheses. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for Social Sciences* (pp. 107-115). Thousand Oaks: Sage.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests.* Retrieved from https://files.eric.ed.gov/fulltext/ED477932.pdf.

Tapia, M., & Marsh, E. (2006). A validation of the emotional intelligence inventory. *Psicothema, 18*, 55-58.

Welch, C., & Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education*, *6*(1), 1-19.

Wetzel, E., & Hell, B. (n.d.). *Differential item functioning in the AIST-R.* [PDF Slides]. Retrieved from http://www.ecpa11.lu.lv/files/Wetzel.pdf.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three Methods using real data. *International Journal of Testing, 9*, 41-59.

Wood, S. W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small.* (Doctoral dissertation). Retrieved from: http://ir.uiowa.edu/etd/1110.

Young, E. L., & Sudweeks, R. R. (2005). Gender differential item functioning in the Multidimensional Self Concept Scale with a sample of early adolescent students. *Measurement and Evaluation in Counseling and Development, 38*(1), 29-43.

Zwick, R., Thayer, D., & Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items* (CRE Board Professional Report No. 93-10P & ETS Research Report 97-05). New Jersey: Educational Testing Service.