International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Evaluating Items' Fairness in a College Admission Test of a Philippine University

Jose Q. Pedrajita

University of the Philippines

Abstract: ***This study looked into differential item functioning in a College Admission Test (CAT) of a Philippine University using the Rasch Model. The CAT covers language proficiency, science, mathematics, and reading comprehension. Logical data analysis was performed to determine which DIF items were biased. Of the 250 items, 94 or 37.6% showed negligible DIF, 93 or 37.2% displayed moderate DIF, and 63 or 25.2% exhibited severe DIF. Of the 63 items with severe DIF, 13 were declared biased. Sources of bias were categorized into two: construct-irrelevant and construct-relevant variance. Revision or replacement of biased items is recommended.***

Keywords: College Admission Test, differential item functioning, differential item functioning analysis, item bias, logical data analysis, Rasch model

1. Introduction

An acceptable assessment is one that is fair and unbiased to all groups of people sharing the same level of proficiency. Test developers have been trying to come up with unbiased tests through differential item functioning (DIF) procedures so that people of the same level of proficiency would have equal chance of performing well in the test items. Examinees of the same ability but belonging to different groups like male or female are expected to show same level of performance on the test, or to an individual item in the test.

Test fairness is a crucial issue in testing. A test that is not fair is a biased test. One of the important considerations in selection and use of any test is that test must not be biased, that is test must be fair to all candidates. A fair test is one being valid for all groups and individuals providing each person with an equal opportunity of demonstrating his/her skills and knowledge relevant to the purpose of the test.

* *Paper Presented at the 2019 Annual Meeting of the American Educational Research Association, Toronto, Ontario, Canada, April 5-9, 2019. This work was funded by the University System Enhanced Creative Work and Research Grant.*

The study analyzed the results of the 2015 College Admission Test (CAT) of a Philippine University. Specifically, it sought answers to the following questions:

(1) Using the Rasch model, which test items display *differential item functioning* between the following matched examinees: *male-female; public general-private; public general-public science; public general-SUC; SUC-private; University-administered-private; and University-administered-SUC* in the following subtests?
   a) Language Proficiency
      - English
      - Filipino
   b) Science
   c) Mathematics
   d) Reading Comprehension
      - English
      - Filipino

(2) What are the causes of DIF as determined in the logical data analysis procedure?

The CAT is part of the admission requirements of the university, administered to graduates of Philippine and foreign high schools. The five hour exam covers language proficiency, science, mathematics, and reading comprehension. CAT questions can be in English or Filipino.

2. Method

The 2015 CAT was subjected to *differential item functioning* analysis using the Rasch model. Thereafter, the identified DIF items were subjected to logical data analysis.

The Rasch model is a psychometric model for analyzing categorical data as a function of the trade-off between (a) the respondent's abilities, attitudes, or personality traits and (b) the item difficulty. The probability of a correct response is expressed as a mathematical function of examinee ability and item characteristics – also known as item characteristic curve (ICC). The ICC is a graph of a statistical model that expresses the probability of a correct response for every ability level of examinees (DeMars, 2010; Hambleton et al., 1991). The ICC graphically represents the regression of the item score on examinees' ability, which is known as the item response function. This function is plotted with the ability level of examinees along the X-axis, against the probability of answering an item correctly on the Y-axis. An examinee's ability is denoted by the Greek letter theta (θ). Probability of correct response, P(θ), range from 0.00 to 1.00. In typical items, this probability is smaller for individuals with low ability than for those with higher ability levels. Consequently, if the probability function P(θ) is plotted against ability level, the result is the typical S-shaped form of the ICC (see Appendix, Figure 1).

Item characteristic functions are compared in two ways. The most direct approach is to compare the parameters that describe the item characteristic curves (ICC). The second approach is to compare the ICCs by evaluating the area between them.

The difference between uniform and non-uniform DIF can be explained graphically. Figure 2 depicts uniform DIF. In uniform DIF the two curves have the same slope but are different in their locations. They are parallel and are visually separated from each other. In this example, the focal group (Group 2) falls below the reference group (Group 1). This shows possible DIF against the focal group which is consistent across all levels of the matching criterion (see Appendix, Figure 2).

In non-uniform DIF the curves have different slopes and cross, while the location of the curve is the same. Figure 3 depicts nonuniform DIF. In this example, there is DIF against the reference group (Group 1) for low levels of the matching criterion, and DIF against the focal group (Group 2) for higher levels of the matching criterion (see Appendix, Figure 3).

The ideal is that there should be little difference between the ICCs of the two groups being compared as illustrated in Figure 4. The area between the ICCs is zero, the ICCs coincide and, hence, no DIF is present (see Appendix, Figure 4).

To test for significance, the obtained Lord's chi-square values were compared to the critical value ($\chi^2 = 3.84$) at $\alpha = .05$ to determine the presence of DIF. Items were flagged as displaying DIF against a particular group when the computed chi-square values were significant. Likewise, the ΔLord values were used to classify the items into three categories to reflect the degree of DIF, labeled A, B, and C, which indicated negligible, moderate, and large DIF respectively. DIF is considered *negligible* when the magnitude of ΔLord < 1 or ΔLord is not statistically different from zero. DIF is considered *moderate* when ΔLord is statistically different from zero and either a) ΔLord < 1.5 or b) ΔLord not significantly different from 1.0. DIF was considered *large* when ΔLord is significantly greater than 1.0 and ΔLord is equal to or greater than 1.5.

The research hypothesis states that *Item response functions are different for the reference-focal matched groups over all items in the Language Proficiency in English and in Filipino, Science, Mathematics, and Reading Comprehension in English and Filipino.*

Table 1 shows the distribution of the examinees by type of high school.

Table 1: Distribution of 2015 CAT Examinees by Type of High School

Type of High School	No. of Examinees	Percent
Public General	25173	29.11
Public Science	6132	7.09
Public Vocational	949	1.10
Public Barrio/Barangay	331	0.38
University-administered	326	0.38
State University/College	2680	3.10
Private	50216	58.07
Middle East	638	0.74
PEPT	7	0.01
Foreign	25	0.03
TOTAL	86477	100.00

he highest number of examinees, 50,216 or 58.07 percent, came from *private* high schools. Examinees from *public general* came second with 25,173 or 29.11 percent. Third, are examinees from *public science* with 6,132 or 7.09 percent. The least number of examinees were those who passed PEPT. Overall, the total number of examinees adds up to 86,477 including foreign applicants.

The number of examinees by gender is presented in Table 2. There were more female than male examinees.

Table 2: Distribution of CAT 2015 Examinees by Gender

Sex	No. of Examinees	Percent
Male	31623	36.57
Female	54854	63.43
TOTAL	86477	100.00

Empirical evidence of differential performance is necessary, but not sufficient to draw conclusion that bias is present (Hambleton (1991). The condition that the item is biased requires a logical analysis in addition to a significant DIF index (Camili and Sheppard, 1994). Logical analysis helps determine why the items seem to be relatively more difficult to a particular group (Wolf & Phyllis, 1990). A common practice of test developers seeking to eliminate "bias" in educational and psychological tests has been to arrange for a panel of experts to review the test items (Reynolds, 2006).

Hence, eight test experts were invited to participate in the FGDs. The composition of the FGDs is shown in Table 3.

Table 3: Participants for the Focus Group Discussion

Department/Institute of Test Experts	No. of Experts	FGD Participated
Department of Psychology	2	All FGDs
Department of English and Comparative Literature	1	LPE
Departamento ng Filipino at Panitikan	2	LPF & RCF
REGALE of the College of Education	1	RCE
Institute of Biology	1	SCIENCE
Institute of Mathematics	1	MATH
Total	8	

Table 4 presents the classification of the admission test items. Language Proficiency-English comprised of 40 items, Language Proficiency-Filipino composed of 40 items, Science consists of 50 items, Mathematics comprised of 50 items, Reading Comprehension-English consists of 35 items, and Reading Comprehension-Filipino is composed of 35 items as well.

Table 4: The 2015 CAT Subtests

Subtest	No. of Items	Item No.
Language Proficiency – English (LPE)	40	1 – 46
Language Proficiency – Filipino (LPF)	40	51 – 90
Science (SCIENCE)	50	91 – 149
Mathematics (MATH)	50	151 – 210
Reading Comprehension – English (RCE)	35	211 – 245
Reading Comprehension – Filipino (RCF)	35	246 – 280

3. Results

Differential Item Functioning Analysis

Gender: The analysis provides the obtained Lord's chi-square values which were compared to the critical value (3.84) either at .05 alpha level with one degree of freedom to determine significance, that is, the presence of DIF. The delta-Lord (ΔLord) values shows the degree of DIF; the item may be free of DIF ("A" item), showing moderate DIF ("B" item), or having severe DIF ("C" item). Items tagged as "A" and "B" was not reflected in the table. Only the "C" items, which are the concerns in the DIF analysis, were reflected. The "C" items are recommended for strict revision or to be replaced. The last column in each table provides the conclusion about the group found to be potentially biased on each item Actual items cannot be presented because of confidentiality reasons and restrictions. The ICCs for the gender-based C items were also reflected for visual inspection of the degree of DIF.

Table 5 presents the DIF items between the male and *female* examinees. Two items in Language Proficiency-English were detected displaying DIF, namely, Items 7 and 44. Both items displayed large magnitude of DIF. Item 7 is potentially biased against the focal group (*female examinees*), while Item 44 is potentially biased against the reference group (*male examinees*). Items tagged as "A" consisted of 37 items and item tagged as "B" consists of only one item (Item 13). Their item characteristic curve is shown in Figure 5.

Table 5: Gender-based DIF Items

Matched Group	Subtest	Item	Rasch DIF Indices				MH DIF Indices				Potentially Biased Group
					Effect Size				Effect Size		
			Lord's X^2	*p*	ΔLord	Code	MHX^2	*p*	ΔMH	Code	
Male-Female	LPE	7	2454.17	.000***	2.13	C	2561.07	.000***	1.97	C	Female
	n = 63,242	44	2150.96	.000***	-1.90	C	1997.93	.000***	-1.66	C	Male
	LPF	------	------	-----	-----	-----	------	-----	-----	------	------
	n = 62,840										
	SCIENCE	------	------	-----	------	-----	------	-----	-----	------	------
	n = 60,512										
	MATH	------	------	-----	------	-----	------	-----	-----	------	------
	n = 60,812										
	RCE	------	------	------	------	------	------	-----	-----	------	------
	n = 63,238										
	RCF	------	------	------	------	------	------	-----	-----	------	------
	n = 63,104										

Note: A = Free of DIF
B = Moderate DIF, Not Ideal Item but Acceptable
C = Severe DIF, Item Needs Replacement
Focal Group = Female
Reference Group = Male
*p < .05 **p < .01 ***p < .001

Language Proficiency-Filipino yielded no "C" item. Items tagged as "A" consisted of 39 items and item tagged as "B" consists of only one item.

Science yielded no items displaying *gender-based* DIF. There were no items tagged as "B" and "C". All items were tagged as "A", indicating negligible DIF.

Like in Science, Mathematics yielded no items displaying *gender-based* DIF. There were no items tagged as "B" and "C". All items were tagged as "A".

In like manner with Science and Mathematics, Reading Comprehension-English yielded no items displaying *gender-based* DIF. Also, there were no items tagged as "B" and "C". All items were tagged as "A".

Reading Comprehension-Filipino yielded no "C" item. A single item (Item 265) was flagged as displaying moderate DIF. Items tagged as "A" consists of 34 items.

Overall, out of 250 items in the 2015 CAT, only two items with large DIF between *male* and *female* examinees were identified. These two items were identified in Language Proficiency-English.

Type of High School. Only the "C" items were reflected in the tables. The last column in each table provides the group which is potentially biased on each item. Actual items cannot be presented because of confidentiality reasons and restrictions.

I. Public General and Private High Schools

Table 6 shows the result of the DIF analysis between examinees from *public general* and *private* high schools.

Table 6: DIF Items in the Public General-Private High Schools Matched Examinees

Matched Group	Subtest	Item	Rasch DIF Indices				MH DIF Indices				Potentially Biased Group
			Lord's X^2	*p*	Effect Size ΔLord	Code	MHX^2	*p*	Effect Size ΔMH	Code	
Public General-Private High Schools	LPE *n = 50,332*	26	1316.98	.000***	- 2.11	C	473.42	.000***	- 1.24	B	Pub Gen
	LPF *n = 49,934*	-----	------	-----	-----	-----	------	------	-----	-----	------
	SCIENCE *n = 50,346*	-----	------	------	-----	-----	------	------	-----	-----	------
	MATH *n = 50,346*	-----	------	------	-----	-----	------	------	-----	-----	------
	RCE *n = 50,332*	-----	------	------	-----	-----	------	------	-----	-----	------
	RCF *n = 50,346*	-----	------	------	-----	-----	------	------	-----	-----	------

Note: A = Free of DIF
B = Moderate DIF, Not Ideal Item but Acceptable
C = Severe DIF, Item Needs Replacement
Focal Group = Private HS
Reference Group = Public General HS
*p < .05 **p < .01 ***p < .001

Language Proficiency-English yielded one "C" item (Item 26). Items tagged as "A" consist of 38 items. Only one item tagged as "B" (Item 1) was identified. Item 26 was flagged as displaying severe DIF against the reference group (*public general*). The ICC of item 26 is shown in Figure 6.

In Language Proficiency-Filipino, no "C" items were detected. However, there were four "B" items, namely, Items 56, 60, 62 and 65. Items tagged as "A" consists of 36 items.

Science yielded no items displaying large magnitude of DIF between examinees from *public general* and *private* high schools. All items were tagged as "A".

Mathematics subtest yielded no items with large DIF between the *public general* and *private* high school examinees. However, a "B" item (Item 164) was detected. The ΔLord effect size reveals that item 164 showed moderate DIF. Items tagged as "A" consisted of 49 items.

In like manner with Science, the Reading Comprehension-English yielded no items displaying large magnitude of DIF between the *public general* and *private* high schools' examinees. All items were tagged as "A".

In the *public general-private* high schools matched examinees, Reading Comprehension-Filipino yielded 34 "A" items, 1 "B" item but no "C" item. The B item is Item 258.

Overall, only one item with severe DIF effect between *public general* and *private* high school examinees was detected. Such item was identified in Language Proficiency-English and is potentially biased against the *public general* high school examinees.

II. Public General and Public Science High Schools

Language Proficiency-English yielded 34 "A" items, four "B" items, and two "C" items. The "C" items were items 1 and 26. Their Lord's chi-square values were significant. Their ΔLord values show that they were displaying severe DIF. Item 1 disadvantaged the focal group (*public science*). While, item 26 disadvantaged the reference group (*public general*). The ICCs for items 1 and 26 is shown in Figure 7a.

Language Proficiency-Filipino incurred 36 "A" items, 3 "B" items, and 1 "C" item. The Lord's chi-square values of the "B" and "C" items were significant. The "B" items were Items 52, 64 and 68. Item 56 was affected by severe DIF and it disadvantaged the reference group (*public general*). The ICC is shown in Figure 7b.

Table 7 presents the DIF analysis results between the *public general* and *public science* high school examinees.

Table 7: DIF Items in the Public General-Public Science Matched Examinees

Matched Group	Subtest	Item	Rasch DIF Indices				MH DIF Indices				Potentially Biased Group
			Lord's X^2	*P*	Effect Size ΔLord	Code	MHX^2	*p*	Effect Size ΔMH	Code	
Public General-Public Science High Schools	LPE *n = 10,576*	1	318.59	.000***	1.54	C	-----	-----	-----	-----	Pub Sci
		26	344.43	.000***.	1.68	C	-----	-----	-----	-----	Pub Gen
	LPF *n = 12,220*	56	489.56	.000***	-1.81	C	515.16	.000***	-1.63	C	Pub Gen
	SCIENCE *n = 10,608*	91	275.46	.000***	-1.75	C	------	-----	-----	-----	Pub Gen
	MATH	177	866.28	.000***	3.27	C	------	-----	-----	-----	Pub Sci

	n = 10,810										
	RCE	227	332.02	.000***	-2.00	C	89.77	.000***	-1.08	B	Pub Gen
	n = 11,600										
	RCF	-----	-----	-----	-----	-----	-----	-----	-----	-----	------
	n = 11,944										

Note: A = Free of DIF
B = Moderate DIF, Not Ideal Item but Acceptable
C = Severe DIF, Item Needs Replacement
Focal Group = Public Science
Reference Group = Public General
*p < .05 **p < .01 ***p < .001

Science subtest incurred 45 "A" items, four "B" items and one "C" item. The Lord's chi-square values of the four "B" items and the "C" item were significant. Likewise, their ΔLord values show that items 110, 137, 144, and 147 were displaying moderate DIF, while Item 91 was displaying large DIF. Item 91 disadvantaged the reference group (*public general*). The ICC is shown in Figure 7c.

Mathematics subtest yielded 44 "A" items, five "B" items and one "C" item. Their obtained Lord's chi-square values were significant. The ΔLord effect size values of items 159, 162, 166, 173, and 189 indicate displaying moderate DIF, while Item 177 was displaying large DIF. Item 177 disadvantaged the focal group (*public science*). Its ICC is shown in Figure 7d.

The DIF analysis between examinees' from *public general* and *public science* high schools in Reading Comprehension-English yielded 31 "A" items, 3 "B" items (216, 226, and 235) and only one "C" item (227). Their Lord's chi-square values were all significant. Item 227 disadvantaged the reference group (*public general*). The ICC is shown in Figure 7e.

In Reading Comprehension-Filipino, thirty-two (32) items were tagged as "A", three (3) items were tagged as "B" and no "C" item was identified. The obtained Lord's chi-square values of items 254, 267, and 270 were all significant and their ΔLord effect size values indicate that they are showing moderate DIF.

Overall, there were six items with severe DIF between *public general* and *public science* high school examinees. These items were identified in Language Proficiency-English, Language Proficiency-Filipino, Science, Mathematics, and Reading Comprehension-English. Four of them were potentially biased against the *public general*, while two were potentially biased against the *public science* high school examinees.

III. Public General and State University/College High Schools

Language Proficiency-English yielded no items displaying severe DIF between examinees from *public general and State university/College*. All items were tagged as "A" items indicating negligible DIF.

Like the Language Proficiency-English, Language Proficiency-Filipino yielded no items displaying moderate and severe DIF between examinees from *public general* and *State university/College* high schools. All items were tagged as "A".

Forty-nine (49) "A" items and only one "B" item, item 128, in Science was identified as exhibiting moderate DIF between examinees from *public general* and *SUC* high schools' examinees. The obtained Lord's chi-square value of item 128 was significant and its ΔLord effect size value reveals that it is showing moderate DIF. No "C" item was identified.

Mathematics yielded forty-nine (49) "A" items, one "B" item, but no "C" item. The obtained Lord's chi-square value of item 164 was significant and the ΔLord effect size reveals that it is showing moderate DIF.

Just like the Language Proficiency-English and Language Proficiency-Filipino, Reading Comprehension-English incurred no items displaying DIF between examinees from *public general* and *State university/College* high schools. All items were tagged as "A".

Likewise, Reading Comprehension-Filipino incurred no items displaying DIF between examinees from *public general* and *State university/College* high schools. All items were tagged as "A".

Overall, the 2015 CAT is to a very large extent fair between the examinees from public general *and* State university/College *high schools.*

IV. State University/College and Private High Schools

In Language Proficiency-English, 39 items were tagged as "A", one item (item 26) was tagged as "B", and no "C" item was identified. The obtained Lord's chi-square value of item 26 was significant and its ΔLord effect size reveals that item is a "B" item.

Language Proficiency-Filipino incurred no items displaying DIF between examinees from *State university/College* and *private* high schools. All items were tagged as "A".

Likewise, Science incurred no items displaying DIF between examinees from *State university/College* and *private* high schools. All items were tagged as "A".

Mathematics also yielded no items displaying DIF between examinees from *State university/College* and *private* high schools. All items were also tagged as "A".

Likewise, Reading Comprehension-English incurred no items displaying DIF between examinees from *State university/College* and *private* high schools. All items were also tagged as "A".

Also, Reading Comprehension-Filipino yielded no items displaying DIF between examinees from *State university/College* and *private* high schools. All 35 items were tagged as "A".

Overall, no items with large DIF effect between examinees from *SUC* and *private* high schools were identified. With this result, the 2015 CAT is to a very large extent fair between the examinees from *State university/College* and *private* high schools.

V. University-administered and Private High Schools

Table 8 shows the DIF items between the *University-administered-private* matched groups.

Language Proficiency-English incurred thirty-two "A" items, five "B" items, and three "C" items. The five "B" items were items 1, 4, 9, 20, and 32 and the "C" items were items 17, 22, and 38. Item 38 disadvantaged the reference group (*University-administered*). Whereas, items 17 and 22 disadvantage the focal group (*private*). Their ICCs are shown in Figure 8a.

Language Proficiency-Filipino incurred twenty-nine "A" items, six "B" items, and five "C" items. The Lord's chi-square values for the B and C items were significant. The ΔLord effect size values reveal that items 53, 63, 74, 79, 82, and 83 were displaying moderate DIF, while Items 52, 61, 64, 68, and 75 were displaying severe DIF. Items 52, 61, 64, 68, and 75 disadvantaged the reference group (*University-administered*). Their ICCs are shown in Figure 8b.

Science incurred thirty-five "A" items, nine "B" items, and six "C" items. The Lord's chi-square values for each of the DIF items were significant. The ΔLord effect size values reveal that items 100, 107, 114, 118, 119, 134, 137, 141, and 145 were exhibiting moderate DIF while Items 91, 110, 129, 136, 144, and 147 were exhibiting severe DIF. Items 91, 136, 144, and 147 disadvantaged the focal group (*private*). Whereas, items 110, and 129 disadvantage the reference group (*University-administered*). Their ICCs are shown in Figure 8c.

Table 8: DIF Items in the University-administered-Private Matched Examinees

Matched Group	Subtest	Item	Rasch DIF Indices				MH DIF Indices				Potentially Biased Group
			Lord's X^2	*p*	Effect Size		MHX^2	*p*	Effect Size		
					ΔLord	Code			ΔMH	Code	
Univ-adm-Private High Schools	LPE *n = 650*	17	21.43	.000***	1.57	C	19.25	.000***	1.46	B	Private
		22	17.31	.000***	1.55	C	9.07	.0I7*	1.17	B	Private
		38	23.79	.000***	-1.56	C					Univ-adm
	LPF *(n = 652)*	52	54.99	.000***	-2.34	C	21.33	.000***.	-1.30	B	Univ-adm
		61	20.94	.000***	-1.69	C					Univ-adm
		64	35.29	.000***	-1.86	C					Univ-adm
		68	62.04	.000***	-2.60	C					Univ-adm
		75	34.93	.000***	-1.88	C					Univ-adm
		79	15.59	.000***	1.45	B	23.46	.000***	1.68	C	Private
	SCIENCE *(n = 652)*	91	25.82	.000***.	3.26	C	5.79	.02*	2.17	C	Private
		110	30.75	.000***	-1.71	C					Univ-adm
		129	24.49	.000***	-1.59	C					Univ-adm
		136	25.10	.000***	1.94	C	11.14	.042*	1.03	B	Private
		144	20.85	.000***	1.97	C	4.21	.040*	1.03	B	Private
		147	19.74	.000***	1.52	C					Private
	MATH *(n = 652)*	160	17.50	.000***	1.54	C	17.55	.0014**	1.43	B	Private
		165	36.94	.000***	-2.06	C					Univ-adm
		166	31.93	.000***	-1.98	C					Univ-adm
		173	22.56	.000***	-1.74	C					Univ-adm
		177	40.37	.000***	-2.68	C					Univ-adm
		181	19.14	.000***	1.60	C					Private
	RCE *(n = 652)*	225	6.02	.01*	1.58	C		.			Private
		226	10.43	.001**	2.22	C					Private
		227	11.92	.006**	1.89	C					Private
	RCF *(n = 652)*	260	19.83	.000***	-1.63	C					Univ-adm
		271	33.40	.000***	1.94	C					Private

Note: A = Free of DIF
B = Moderate DIF, Not Ideal Item but Acceptable
C = Severe DIF, Item Needs Replacement
Focal Group = Private HS Reference Group = University-administered HS
*p < .05 **p < .01 ***p < .001

Mathematics yielded forty "A" items, four "B" items, and six "C" items. The chi-square values for the four "B" items and six "C" items were significant and their ΔLord effect size values indicate that four items, namely, 167, 171, 176, and 189 showed moderate DIF, while, six items, 160, 165, 166, 173, 177, and 181 exhibit severe DIF.

Items 160 and 181 disadvantaged the focal group (*private*). Whereas, items 165, 166, 173, and 177 disadvantage the reference group (*University-administered*). Their ICCs are shown in Figure 8d.

Reading Comprehension-English yielded twenty-eight "A" items, four "B" items, and three "C" items. The Lord's chi-square values for the B and C items were significant. The ΔLord effect size values of items 216, 231, 235, and 238 indicates moderate DIF, while three items, 225, 226, and 227 were exhibiting large DIF. Items 225, 226, 227 disadvantaged the focal group (*private*). The ICCs are shown in Figure 8e.

Reading Comprehension-Filipino yielded thirty-two "A" items, one "B" item, and two "C" items. The Lord's chi-square values of the B and C items were significant. The ΔLord effect size values reveal that of the three items, item 254 displayed moderate DIF, while items 260 and 271 displayed large DIF. Item 260 disadvantaged the reference group (*University-administered*). Whereas, item 271 disadvantage the focal group (*private*). The ICCs are shown in Figure 8f.

Overall, there were 26 items with large DIF effect between examinees from *University-administered* and *private* high school across the subtests.

VI. University-administered and State University/College High Schools

Language Proficiency-English yielded thirty-one "A" items, five "B" items, and four "C" items. The Lord's chi-square values of the B and C items were significant. The ΔLord effect size values reveal that items 4, 12, 20, 22, and 32 were displaying moderate DIF and items 1, 17, 26, and 38 were displaying large DIF. Items 1 and 38 disadvantaged the reference group (*University-administered*). While, items 17 and 26 disadvantaged the focal group (*SUC*). The ICCs are shown in Figure 9a.

Language Proficiency-Filipino incurred twenty-seven "A" items, five "B" items, and eight "C" items. The Lord's chi-square values of the "B" and "C" items were significant. Their ΔLord effect size values reveal that items 53, 74, 86, 87 and 88 exhibits moderate DIF while items 52, 56, 57, 61, 64, 68, 75, and 79 exhibit large DIF. Items 52, 57, 61, 64, 68, and 75 disadvantaged the reference group (*University-administered*). While, items 56 and 79 disadvantaged the focal group (*SUC*). The ICCs are shown in Figure 9b.

Table 9 shows the results of the DIF analysis between examinees from *University-administered* and *SUC* high schools.

Table 9: DIF Items in the University-administered-SUC Matched Examinees

Matched Group	Subtest	Item	Rasch DIF Indices				MH DIF Indices				Potentially Biased Group
					Effect Size				Effect Size		
			Lord's X^2	*p*	ΔLord	Code	MHX^2	*p*	ΔMH	Code	
Univ-adm-SUC High Schools	LPE *(n = 600)*	1	27.68	.0000***	-1.89	C	14.25	.0064**	-1.34	B	Univ-adm
		17	21.03	.0000***	1.63	C	11.15	.0089**	1.35	B	SUC
		26	22.26	.0000***	1.60	C					SUC
		38	25.07	.0000***	-1.70	C	8.91	.0226*	-1.03	B	Univ-adm
	LPF *(n = 646)*	52	51.41	.0000***	-2.40	C	18.02	.0002***.	-1.37	B	Univ-adm
		56	22.90	.0000***	1.60	C	27.27	.0000***	1.72	C	SUC
		57	22.71	.0000***	-1.61	C	11.55	.0045**	-1.09	B	Univ-adm
		61	15.67	.0004***	-1.54	C					Univ-adm
		64	39.20	.0000***	-2.08	C					Univ-adm
		68	78.88	.0000***	-3.14	C	25.38	.0000***	-1.69	C	Univ-adm
		75	40.96	.0000***	-2.15	C	13.66	.0017**	-1.26	B	Univ-adm
		79	18.60	.0001***	1.65	C	25.41	.0000***	1.85	C	SUC
	SCIENCE *(n = 592)*	91	26.16	.0000***.	3.34	C	9.55	.0020*	2.71	C	SUC
		110	22.76	.0000***	-1.55	C					Univ-adm
		136	20.48	.0001***	1.82	C					SUC
		144	26.51	.0000***	2.29	C	8.61	.0033*	1.44	B	SUC
		147	21.25	.0001***	1.68	C					SUC
	MATH *(n = 594)*	160	17.15	.0002***	1.59	C	18.00	.0011**	1.59	C	SUC
		165	26.39	.0000***	-1.84	C					Univ-adm
		166	25.63	.0000***	-1.87	C					Univ-adm
		173	21.38	.0000***	-1.79	C					Univ-adm
		177	42.16	.0000***	-2.88	C					Univ-adm
		181	17.77	.0002***	1.62	C	8.46	.0453*	1.06	B	SUC
		189	20.76	.0001***	1.59	C					SUC
	RCE *(n =648)*	226	7.02	.0353*	1.86	C		.			SUC
		227	8.04	.0266*	1.59	C					SUC

		238	25.03	.0000***	-1.67	C					Univ-adm
	RCF	254	32.74	.0000***	-2.03	C	15.80	.0012**	-1.42	B	Univ-adm
	(n = 610)	271	34.55	.0000***	2.07	C	24.01	.0000***	1.65	C	SUC

Note: A = Free of DIF
B = Moderate DIF, Not Ideal Item but Acceptable
C = Severe DIF, Item Needs Replacement
Focal Group = SUC HS
Reference Group = University-administered HS
*p < .05 **p < .01 ***p < .001

Science incurred thirty-four "A" items, eleven "B" items, and five "C" items. The Lord's chi-square values for each of the "B" and "C" items were significant. And their ΔLord effect size values reveal that items 105, 106, 114, 118, 119, 125, 129, 134, 137, 141, and 145 showed moderate DIF, while items 91, 110, 136, 144, and 147 showed large DIF. Items 91, 136, 144, and 147 disadvantaged the focal group (*SUC*). Whereas, item 110 disadvantage the reference group (*University-administered*). The ICCs are shown in Figure 9c.

Mathematics incurred 40 "A" items, 3 "B" items, and 7 "C" items. The chi-square values for the B and C items were significant. Their ΔLord effect size values reveal that items 167, 168, and 171 showed moderate DIF, while items 160, 165, 166, 173, 177, 181 and 189 showed large DIF. Items 165, 166, 173, and 177 disadvantaged the reference group (*University-administered)*. While items 160, 181, and 189 disadvantage the focal group (*SUC*). The ICCs are shown in Figure 9d.

Reading Comprehension-English yielded twenty-eight "A" items, four "B" items, and three "C" items. The Lord's chi-square values obtained for the seven items were significant. Likewise, the ΔLord effect size values reveal that items 216, 231, 235, and 244 showed moderate DIF, while items 226, 227, and 238 showed large DIF. Items 226 and 227 disadvantaged the focal group (*SUC*). Whereas, item 238 disadvantage the reference group (*University-administered*). The ICCs are shown in Figure 9e.

Reading Comprehension-Filipino incurred thirty "A" items, three "B" items, and two "C" items. The chi-square values for the DIF items were significant and their ΔLord effect size values reveal that items 258, 270, and 275 exhibited moderate DIF, while items 254 and 271 displayed large DIF. Items 254 disadvantaged the reference group (*University-administered*), while item 271 disadvantaged the focal group (*SUC*). The ICCs are shown in Figure 9f.

Overall, there were 29 items with large DIF effect between examinees from *University-administered* and *SUC* high schools. Fifteen (15) items were potentially biased against the *University-administered* high school examinees, while fourteen (14) items were potentially biased against the *SUC* high school examinees.

Thus, the research hypothesis which states that *item response functions are different between the reference-focal matched groups over all items* was partially confirmed in the DIF items in each subtest.

Table 10 shows the summary of DIF items.

Table 10: Summary of DIF Items

Subtest	Matched Group	DIF Items
LPE	Male-Female	7, 44
	Public General –Private	26
	Public General–Public Science	1, 26
	Public General – SUC	-----
	SUC – Private	-----
	University-administered–Private	17, 22, 38
	University–administered–SUC	1, 17, 26, 38
LPF	Male-Female	-----
	Public General-Private	-----
	Public General-Public Science	56
	Public General-SUC	-----
	SUC-Private	-----
	University-administered-Private	52, 61, 64, 68, 75, 79
	University-administered-SUC	52, 56, 57, 61, 64, 68, 75, 79
Science	Male-Female	-----
	Public General-Private	-----
	Public General-Public Science	91, 144
	Public General-SUC	-----
	SUC-Private	-----
	University-administered-Private	91, 110, 129, 136, 144, 147
	University-administered-SUC	91, 110, 136, 144, 147
Math	Male-Female	-----
	Public General-Private	-----
	Public General-Public Science	177
	Public General-SUC	-----
	SUC-Private	-----
	University-administered-Private	160, 165, 166, 173, 177, 181
	University-administered-SUC	160, 165, 166, 173, 177 181, 189
RCE	Male-Female	-----
	Public General-Private	-----
	Public General-Public Science	227
	Public General-SUC	-----
	SUC-Private	-----
	University-administered-Private	225, 226, 227
	University-administered-SUC	226, 227, 238
RCF	Male-Female	-----
	Public General-Private	-----
	Public General-Public Science	-----
	Public General-SUC	-----
	SUC-Private	-----
	University-administered-Private	260, 271
	University-administered-SUC	254, 271

Logical Data Analysis of DIF Items

For the LDA, six FGDs were conducted (1 each for Science and Mathematics, 2 for Language Proficiency (1 in English and 1 for Filipino), and 2 for Reading Comprehension (1 in English and 1 in Filipino). The two experts from the Department of Psychology participated in all FGDs and the other 6 experts each joined the FGD for each of the subtests. So, there were three (3) experts, which served as item reviewers, for each FGD.

An *Item Review Tool for Potentially Biased Items* containing the items that display large magnitude of differential item functioning was used in the FGD by the test experts. The subtest to which each item belongs and the direction of possible bias are indicated, respectively. The test experts are tasked to carefully examine each item and specify whether or not he/she agrees that the item is *biased against* the group indicated. An item is biased if it *disadvantaged* the group specified. Among the items that the test experts considered as biased, he/she is tasked to kindly write the reason(s) why he/she think that the item is *biased against* the group specified.

Causes of DIF was categorized into *construct irrelevant variance* (CIV) and the *opportunity to learn* (OL). Construct Irrelevant Variance refers to the possible cause of DIF in which one group might do better in the test item not because of the ability on the subject matter, but because they happen to be familiar with the topic or clues in the item. Opportunity to Learn means that because of different curriculum between or among groups, an item might exhibit DIF. In that case, differential student experience in terms of the curriculum is not relevant to the test.

The biased item/s by subtest and matched group, sources of bias and the biased group is shown in Table 11.

Table 11: Biased Item/s by Subtest and Matched Group, Sources of Bias, and the Biased Group

Subtest	Matched Group	Biased Items	Source of Bias	Biased Group
LPE	Male-Female		----------	------
	Public General –Private	26	Poor construction of item and choices	Public General
	Public General-Public Science	26	Poor construction of item and choices	Public General
	Public General – SUC	------	----------	------
	SUC – Private	------	----------	------
	University-administered– Private	38	Poor construction of item and choices	University-administered
	University–administered– SUC	26	Poor construction of item and choices	SUC
		38	Sentence and options not well constructed,	University-administered
LPF	Male-Female	------	----------	------
	Public General-Private	------	----------	------
	Public General-Public Science	56	Multi-dimensional	Public General
	Public General-SUC	------	----------	------
	SUC-Private	------	----------	------
	University-administered-Private	61	Multi-dimensional, poor item construction	University-administered
	University-administered-SUC	56	Multi-dimensional, poor item construction	SUC
		57	Multi-dimensional, poor item construction	University-administered
		61, 64	Multi-dimensional	University-administered
Science	Male-Female	------	----------	------
	Public General-Private	------	----------	------
	Public General-Public Science		----------	------
	Public General-SUC	------	----------	------
	SUC-Private	------	----------	------
	University-administered-Private	110	Poor item construction	University-administered
	University-administered-SUC	110	Poor construction of item and choices	University-administered
Math	Male-Female	------	----------	------
	Public General-Private	------	----------	------
	Public General-Public Science	177	Confusing construction of the item problem	Public Science
	Public General-SUC	------	----------	------
	SUC-Private	------	----------	------
	University-administered-Private	177	Confusing construction of the item problem	University-administered
	University-administered-SUC	177	Poor item construction	University-administered
RCE	Male-Female	------	----------	------
	Public General-Private	------	----------	------
	Public General-Public Science	227	Perspective presented in the text not consistent with construct	Public General
	Public General-SUC	------	----------	------
	SUC-Private	------	----------	------
	University-administered-Private	225, 227	Poor construction of item	Private
		226	Skill classification is inconsistent with the construct/Poor item construction	Private

	University-administered-SUC	226, 227	Poor item construction/Not consistent with the construct	SUC
		238	Poor item construction/Not consistent with the construct	University-administered
RCF	Male-Female	------	----------	------
	Public General-Private	------	----------	------
	Public General-Public Science	------	----------	------
	Public General-SUC	------	----------	------
	SUC-Private	------	----------	------
	University-administered-Private	260	Poor construction of item and choices	University-administered
	University-administered-SUC	------	----------	------

Language Proficiency-English. Three comparison groups, *male-female, public general-SUC, and SUC-private,* have not incurred biased items, while the other four reference-focal comparison groups had.

Item 26 was declared biased in the following reference-focal group comparisons, *public general-private, public general-public science, and university-administered-SUC*. While, item 38 was declared biased in the *university-administered-private,* and *university-administered-SUC* matched examinees.

In Item 26 the examinees were tasked to choose from the given options the correct mood of the verb to be used in order to make the sentence grammatically correct. The item stem has a blank where the correct mood of the verb should be placed. It measures skill in content words. The item is difficult (p = 0.26) but it has an excellent discrimination index (D = 0.62).

The reviewers unanimously judged item 26 as biased against the reference group (*public general*) in both *public general-private* and *public general-public science* matched group. Likewise, in the *university-administered-SUC* matched group item 26 was considered biased against the focal group (*SUC*). Their comments, common in the three matched groups, were: *"The option are poorly constructed"(R1); "Poorly constructed"(R2); and "Bias may come from the item being difficult in general; or bias comes from poor construction of items and choices."(R3.)* Thus, the source of bias for item 26 is poor construction of the item and choices. Such source of bias is construct-irrelevant.

Item 38 is about cohesion/coherence in sentence order which measure skill in sentence sense and structure. This item requires the examinees to arrange the statements/sentences in correct order. This item seems difficult (p = 0.30) but has a good discrimination index (0.33). Item 38 was regarded biased against the reference group (*university-administered*) in both *university-administered-private* and *university-administered-SUC* matched group since the sentence is not well constructed as well as the sequence of options. The reviewers' comments were: *"Sentence order not well constructed"(R1); "Sentence not well constructed"(R2); and "Problem in sentence order; sequence of options"(R3)* Poor sentence construction and sequence of options is construct-irrelevant, hence, item 38 was biased against the reference group (*University-administered*).

Language Proficiency-Filipino. No biased item emerged in the following matched groups: *male-female, public general-private, public general-SUC, and SUC-private*. While, biased items were detected in the *public general-public science, university-administered-private,* and *university-administered-SUC* matched groups.

Item 56 was declared biased against the reference group (*public general*) in the *public general-public sci*ence matched examinees as well as against the focal group in the *university-administered-SUC* matched group. Item 56 is like a word analogy item and is written in Filipino language. It is about *pagtukoy at pagsusuri sa mga pagbabagong morpoponemiko ng mga salita* (identifying and analyzing morpho-phonemic changes in words) which measures skill in *Estruktura at Kahulugan ng Salita* (structure and meaning of words). Such item is quite difficult (p = 0.31) and has an excellent discrimination index (D = 0.41).

Item 57 was decided biased against the reference group (*university-administered*) in the university-administered-SUC matched group. Item 57 is about pagkilala at pag-iiba sa mga salita ayon sa kasingkahulugan o kasalungat na kahulugan (word analogy) which falls under the subtest skill Estruktura at Kahulugan ng Salita (structure and meaning of words). Item 57 is written in Filipino language. It is similar to a word analogy item which consists of either pair of synonym, pair of antonym, or pair of synonym-antonym. The item requires the examinees to choose which of the given options has the same or opposite meaning as that of the given pair in the item stem. Such item is of optimum difficulty (p = 0.55) and with good discrimination index (D = 0.32).

Both items 56 and 57 have been found to have two sources of bias, multi-dimensionality and poor item construction. That is, it measures more than one skill which is construct-irrelevant. The reviewers' comments are: *"measures more than one skill; poor item construction"(R1); "also measures analogical task"(R2);and "needs to use logic to be able to get the skill needed"(R3).*

Item 61 and item 64 was found biased against the reference group (*university-administered*) in both the *university-administered-private* and the *university-administered-SUC* matched groups. Item 61 is about the choice of the correct pair of singular-plural form of panghalip (pronoun) that is equivalent to the given pair of pronoun in the item stem. The item measures skill in salitang pangnilalaman (parts of speech). It is an easy item (p = 0.77) with fair discrimination

index (D = 0.22). While, Item 64 is about the choice of the correct noun (pangngalan), to be written on the blank provided in the item stem, that will make the sentence correct. It also measures skill in salitang pangnilalaman. It has an optimum difficulty index (p = 0.61) and with a fair discrimination index (D = 0.25).

The result of the FGD points to poor item construction and multidimensionality as the sources of bias. Specifically, for item 61 the reviewers' comments are as follows: *"measures more than one skill; poor item construction"(R1); "measures analogical task"(R2);* and *"knowledge + skill + logic"(R3).* These sources of bias are construct-irrelevant. Therefore, item 61 was biased against the reference group (*university-administered*).

Item 64 was decided to have these sources of bias, *multi-dimensionality* and *sampling* and *curriculum difference*. For this item, the reviewers came up with these comments: *"sampling and curriculum difference"(R1); "also measured analogical task"(R2*)**;** and **"***needs to use logic to be able to get the skill needed"(R3).* Except for sampling and curriculum difference, multi-dimensionality is construct-irrelevant. Eventually, the reviewers decided that items 61 and 64 are both biased against the reference group (*university-administered*).

*Science***.** Five comparison groups, *male-female, public general-private, public general-public science, public general-SUC, and SUC-private,* were not affected by biased items, while the other two reference-focal comparison groups, namely, *university-administered-private*, and *university-administered-SUC* have been affected.

Item 110 was found biased against the reference group (*university-administered*) both in the *university-administered-private* as well as in the *university-administered-SUC* matched groups. Item 110 is about patterns of reproduction in Biology. The item state that the reproductive activities of almost all animals are cyclic and it ask which one of the given options is the best reason for this seasonality. The item has an optimum difficulty level (p = 0.43) but with fair discriminating power (D = 0.21). Specifically, the reviewers came up with these comments: *"Item choices may be too long"(R1); "Options are too long"(R2); and "The item appears to be a reading comprehension type as well as a science question"(R3).*

*Mathematics***.** Four comparison groups, *male-female, public general-private, public general-SUC, and SUC-private,* were not affected by biased items, while the other three reference-focal comparison groups, namely, *public general-public science, university-administered-private,* and *university-administered-SUC* have been affected.

Item 177 was found biased against the focal group (*public science*) in the *public general-public science* matched group, and against the reference group (*university-administered*) in both the *university-administered-private* as well as *university-administered-SUC* matched groups. Item 177 is a problem solving item in Algebra. It is a word problem involving linear and/or quadratic. It measures skill in Algebra. It is a difficult item (p = 0.17) and it does not discriminates well between the high and the low scorers (D = 0.02). The reviewers gave the following comments: *"Confusing construction of the item problem"(R1 & R2); "Confusing construct of the problem, which might have also affected their comprehension"(R3).*

Reading Comprehension-English. Four comparison groups, *male-female, public general-private, public general-SUC, and SUC-private,* were not affected by biased items, while the other three reference-focal comparison groups, namely, *public general-public science, university-administered-private*, and *university-administered-SUC* have been affected.

In Reading Comprehension-English, four DIF items were detected, namely, items 225, 226, 227, and 238. Items 225, 226, and 227 have common two-paragraph narrative but with different item stem each.

In Item 225 the examinees should read the two-paragraph article about play as an issue in early childhood education prior to answering the question in the item stem. It asked "if preschooler play seems to be more social in nature than the play of infants," then what conclusion can be deduced from the given narrative. The examinees will choose from the given options the correct conclusion. This item measures skill in interpretation. This is an easy item (p = 0.85) with fair discrimination index (D = 0.28). Item 225 obtain the following comments: *"poorly constructed item"(R1 & 2); and "perspective presented in the text (that of preschool teachers); poorly constructed item"(R3).*

While, in Item 226 the examinees should read the same two-paragraph article about play as an issue in early childhood education prior to answering the question in the item stem. It asked "what is the message of the article to preschool teachers?" then the examinees will choose from the given options the correct message conveyed by the article. This item is about expressing ideas in alternative ways which measures the subtest skill *creative*. This is also an easy item (p = 0.82) with an excellent discrimination index (D = 0.42). The reviewers found Item 226 biased against the focal group (private) in the *university-administered-private* matched group due to "*skill classification is inconsistent with the construct and poor item construction";* and also found biased against the focal group (*SUC*) in the *university-administered-SUC* matched group.

In Item 227 the examinees should read the same two-paragraph article about the play as an issue in early childhood education prior to answering the question in the item stem. It asked "If preschool teachers would prefer the usual school activity over play in the classroom, children would…." then the examinees will choose from the given options the correct cause and effect relationships or problem-solutions to the issue conveyed by the article. This item is about determining cause and effect relationships or problem-solutions which measures the subtest skill *interpretation*. This is also an easy item (p = 0.75) with an excellent discrimination index (D = 0.56). Item 227 was regarded as biased against the reference group (*public general*) in the *public general-public science* matched group due to "*perspective presented in the text not consistent with*

construct."; biased against the focal group (*private)* in the *university-administered-private* matched group due to "*poor item construction."*; and biased against the focal group (*SUC*) in the *university-administered-SUC* matched group due to *"poor item construction and not consistent with the construct."*

In item 238 the examinees has to read the poem prior to answering the question in the item stem. It asked "In the first stanza, the poet creates the image of a war that. . . ," then the examinees will choose the correct idiomatic expression that correctly reflect the image of a war from the given options. The item is about identifying idiomatic expressions which measure the subtest skill *critical evaluation*. The item is of optimum difficulty (p = 0.36) and with a fair discrimination index (D = 0.21). The reviewers found Item 238 biased against the reference group (*university-administered*)) in the *university-administered-SUC* matched group due to *"poor construction of choices"(R1); "biased – not rooted to construct"(R2); and "poor item construction, particularly writing of options"(R3).*

*Reading Comprehension-Filipino***.** Six comparison groups, *male-female, public general-private, public general-public science, public general-SUC, SUC-private* and *University-administered-SUC* have not incurred biased items, while the lone reference-focal comparison group, the *university-administered-private*, is affected.

In Item 260 the examinees is tasked to view the two pictures of an Igorot and read the two-paragraph article (written in Filipino) about the metamorphosis of such Igorot under the educational system of the American colonial government in the Philippines prior to answering the question in the item stem. It asked "Ano ang ipinapalagay ng pariralang, "sinadyang piliin at pagtabihin…"? (What does the phrase 'intentionally chosen and analyze' imply…?) then the examinees will choose from the given options the correct message conveyed by the given phrase (parirala). This item is about *pagkilala sa salita, parirala, at pangungusap na ginamit nang iba sa karaniwang konteksto* (word, phrase, and sentence recognition used in context) which measure the subtest skill *mapanuring pag-unaw*a (critical thinking). This is a very difficult item (p = 0.17) with a fair discrimination index (D = 0.15). Item 260 got these comments: *"poor construction of the stem question"(R1); "poor item construction, stem and options"(R2); and "problem on stem; construction of question"(R3).*

Summary of Biased Items. The *University-administered-SUC* matched group obtained the highest number of biased items which is 11, followed by the *University-administered-private* matched group with 8 biased items, then by the *public general-public science* matched group with 4 biased items, and the matched group with the least number of biased items is the *public general-private* with only one biased item. By match groups, the total number of biased items adds up to 24.

However, since some of the items were judged as biased two and/or three times in two or three comparison groups the actual number or identity of biased items across subtests were only thirteen, namely, items 26, 38, 56, 57, 61, 64, 110, 177, 225, 226, 227, 238, and 260.

Four comparison groups, namely, *public general-private*, *public general-public science*, *University-administered-private*, and *University-administered-SUC* incurred biased items. While three comparison groups have not, namely, the *male-female*, *public general-SUC*, and the *SUC-private.*

4. Conclusions

Based on the findings, the following conclusions were arrived at:

Four comparison groups were affected by biased items, namely, *public general-private*, *public general-public science*, *University-administered-private*, and *University-administered-SUC.* Whereas, the *male-female*, *public general-SUC*, and the *SUC-private* were not. Two comparison groups, the *public general-SUC* and *SUC-private*, incurred no DIF and biased items.

The test items were to a very large extent fair between the *male-female, public general-SUC,* and *SUC-private* matched examinees which incurred no biased items; to a large extent fair between public-general-private which incurred 1 biased item; generally fair between the *University-administered* and *private* examinees which incurred 8 biased items and between the *University-administered-SUC* comparison group which incurred 11 biased items.

Of the 250 items in the College Admission Test, 94 or 37.6% showed negligible DIF, 93 or 37.2% displayed moderate DIF, and 63 or 25.2% exhibited severe DIF. Of the 63 items with severe DIF, 13 were declared biased.

A nearly equal number and percentage of DIF items which disadvantage the focal group (83 or 33.2%) and the reference group (84 or 33.6%) across subtests was evident.

Construct-irrelevant and *construct-relevant sources of bias emerged from the findings.* The *construct-irrelevant* sources of bias were: a) poor item and choices construction, and b) multi-dimensionality. The *construct-relevant* sources of bias were: a) opportunity to learn, that is, difference in group exposure to concept reflected in the item, and b) sampling size bias.

5. Recommendations

Based on the findings and conclusions the following recommendations were arrived at:

This study focused only on 7 matched groups of examinees out of 21 possible comparison groups. DIF analysis of the other comparison groups should be conducted to have a clearer and whole picture of the fairness of the items.

DIF analysis of the other editions of the CAT should be conducted. It could be useful for future research work to extend the present research in this direction.

The 2015 CAT must be expurgated of biased items. Biased items should be revised or replaced so as to maintain if not improved the content and reliability of the admission test. By eliminating, replacing, or revising these biased items a more valid, reliable, and fair test would be made.

Items exhibiting DIF be further reviewed by curriculum specialists before further use. The impact of high occurrence of DIF needs further investigation.

Educational institutions, assessment experts, and test developers should consider giving increasing attention to equity of test scores for various groups or subgroups of examinees. Test equity can be achieved by ensuring that a test measures only construct-relevant differences between groups or subgroups of examinees. To achieve test equity among groups or subgroups of examinees, DIF/bias detection must be conducted especially for very important tests like entrance examination and professional licensure examination.

References

[1] Baghi, H. & Ferrara, S. F. (1990). Detecting Differential Item Functioning Using IRT and Mantel-Haenszel Techniques: Implementing Procedures and Comparing Results. *Paper Presented at the Annual Meeting of the Eastern Educational Research Association*, Clearwater, FL, February 14-17, 1990. Retrieved November 2011 from ProQuest Journals.

[2] Camilli and Shepard (1994). Methods for Identifying Biased Test Items. London: SAGE Publications, Inc., California.

[3] Demars, C. (2010). Item Response Theory. New York: Oxford University Press Inc.

[4] Doolittle, A. & Welch, C. (1989). Gender Differences in Performance on a College-level Achievement Test. *ACT Research Report Series*.

[5] Edelen, M. O. & Reeve, B. (2007). Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. *Quality Life Research*. 16: 5-18.

[6] Ercikan, K. & Mendes-Barnett S. (2006). Examining Sources of Gender DIF in Mathematics Assessments Using a Confirmatory Multidimensional Model Approach. *Applied Measurement in Education*. 19(4), 289-304.

[7] Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. London: Lawrence Erlbaum Associates, Publishers.

[8] Gierl, M. J. (1999). *Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications.* Retrieved March 2, 2010 from http://www2.education.ualberta.ca/educ/psych/crame/files/ dif_csse99.pdf

[9] Gruijter, D. M. & Kamp, L. J. (2008). Statistical Test Theory for the Behavioral Sciences. New York: Taylor & Francis Group.

[10] Hambleton, R.K., Swaminathan, H., & Rogers, J.H. (1991). Fundamentals of Item Response Theory (IRT). London: Sage Publications.

[11] Kanjee, Anil (2007). Using Logistic Regression to Detect Bias when Multiple Groups are Tested. *South Africa Journal of Psychology*, 37(1), pp. 47-61.

[12] Koretz, D. (2008). Measuring Up: What Educational Testing Really Tells Us. USA: Harvard University Press.

[13] Le, V. (1999). Identifying Differential Item Functioning on the NELS:88 History Achievement Test. *Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.* Retrieved November 2011 from ProQuest Journals.

[14] Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment.* Vol. 1, No. 1, pp. 1-11.

[15] Osterlind, S. J. (1983). Test Item Bias Detection. London: Sage Publications.Inc., California.

[16] Reynolds R.C., Livingston, R.B., & Willson, V. (2006). Measurement and Assessment in Education. Boston: Pearson.

[17] Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing*. Retrieved Octiber 1, 2010, from the http://www2.hawaii.edu/~roever/brownbag.pdf

[18] Roussos, L. A. & Stout, W. F. (2000). *A Formulation of the Mantel-Haenszel Differential Item Functioning Parameter with Practical Implications.* Retrieved March 2, 2010 fromhttp://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_ storage_01/0000019b/80/1a/70/b1.pdf

[19] Sheppard, R., Han, K., Colarelll, S. M., Dai, G., & King, D. W. (2006*). Differential Item Functioning by Sex and Race in the Hogan Personality Inventory.* SAGE Publications. Retrieved June 1, 2010 from ProQuest Educational Journals

[20] Stoneberg, B.D. (2004). A Study of Gender-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenzel Chi Square Test. *The University of Maryland Measurement, Statistics, and Evaluation Department and the National Center for Education Statistics (NCES) Assessment Division.*

[21] Welch, C. J. & Doolittle, A. E. (1989). Gender-Based Differential Item Performance in English Usage Items. *ACT Research Report Series*.

[22] Wolf, L. R. & Phyllis, R. (1990). The SAT Gender Gap. *Women and Languge*. Vol. 13, Issue 2. Retrieved September 26, 2011 from ProQuest Journal

Appendix

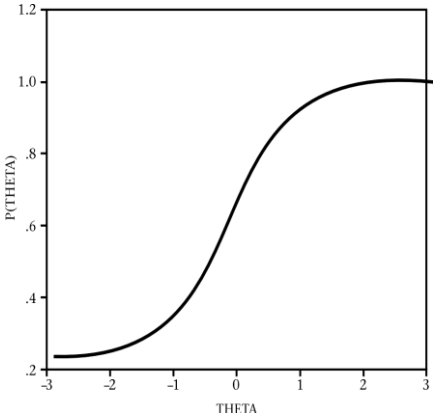

Figure 1: Example of an Item Characteristic Curve (ICC)

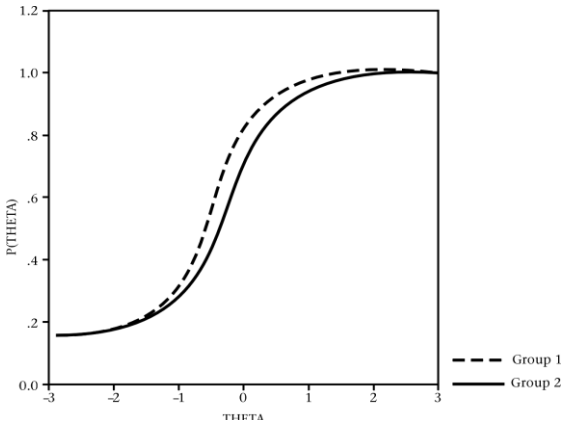

Figure 2: ICC for Uniform DIF

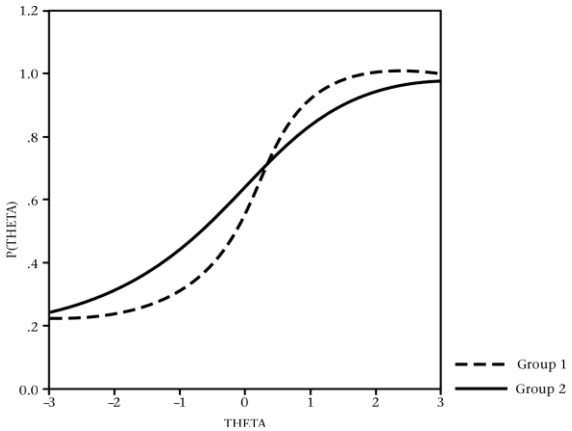

Figure 3: ICC for Non-Uniform DIF

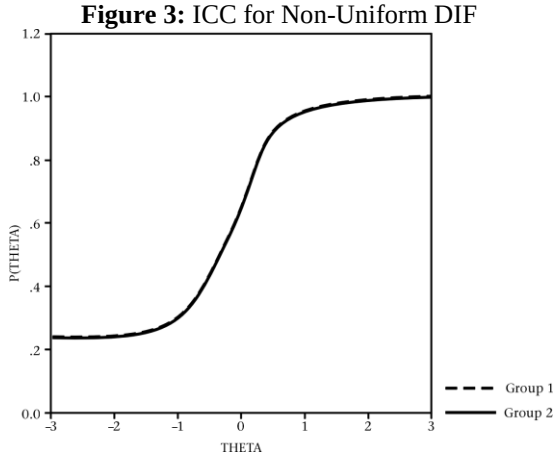

Figure 4: ICC of no DIF

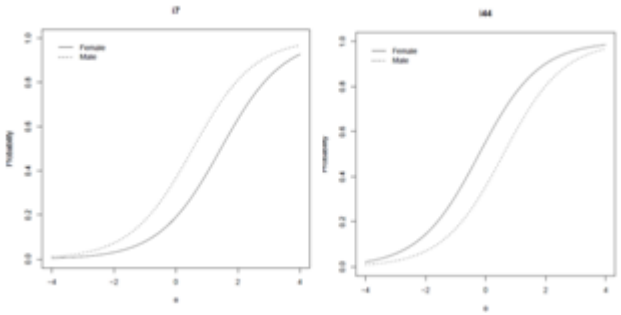

Figure 5: ICCs of Items 7 and 44

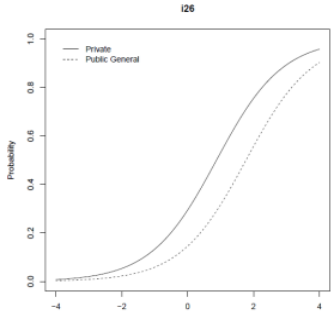

Figure 6: ICC of Item 26

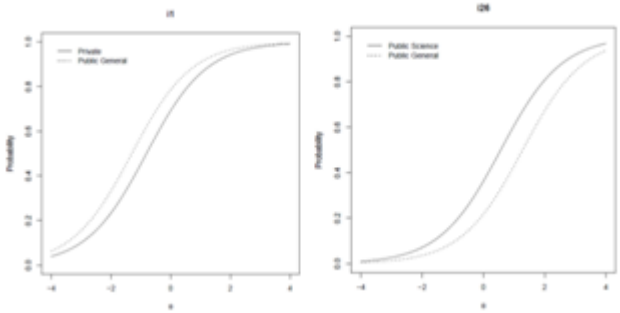

Figure 7(a): ICCs of Items 1 and 26

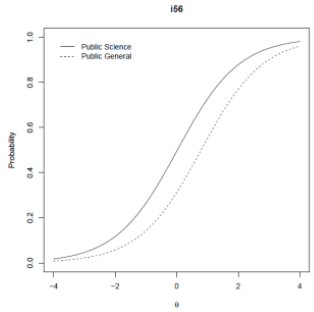

Figure 7(b): ICC of Item 56

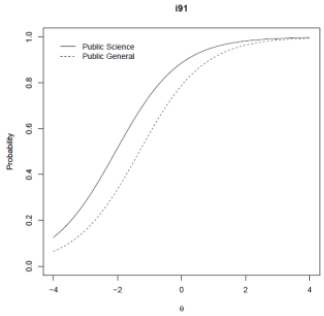

Figure 7(c): ICC of Item 91

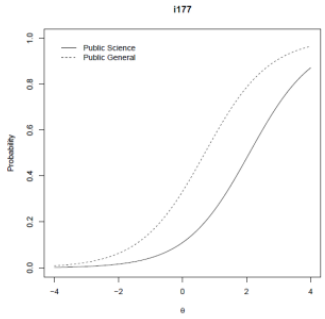

Figure 7(d): ICC of Item 177

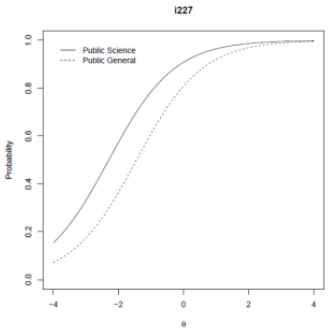

Figure 7(e): ICC of Item 227

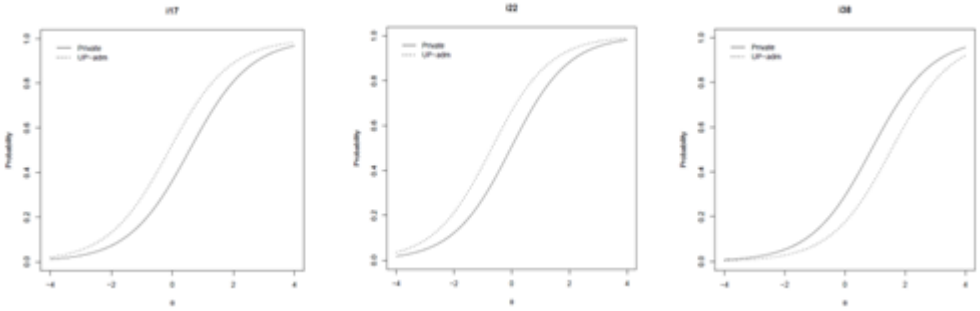

Figure 8 (a): ICCs of Items 17, 22, and 38

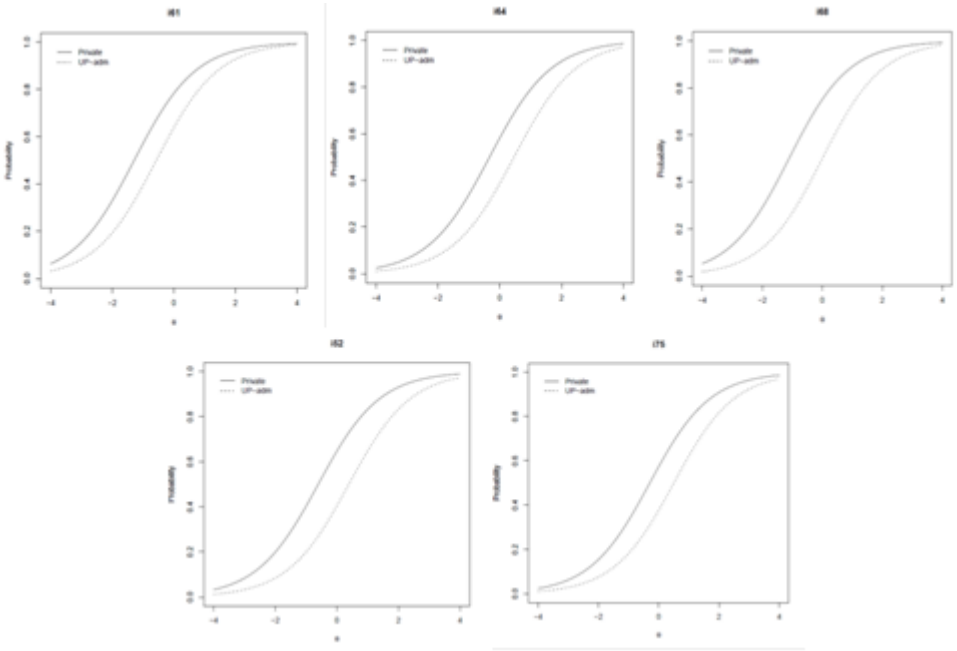

Figure 8(b): ICCs of Items 52, 61, 64, 68, and 75

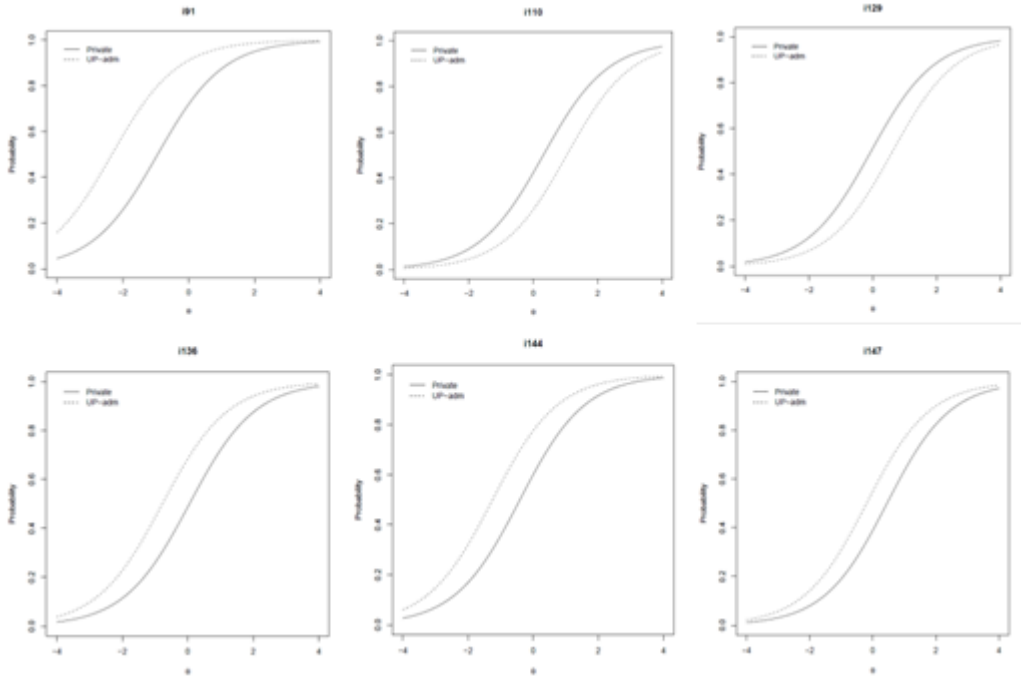

Figure 8 (c): ICCs of Items 91, 110, 129, 136, 144, and 14

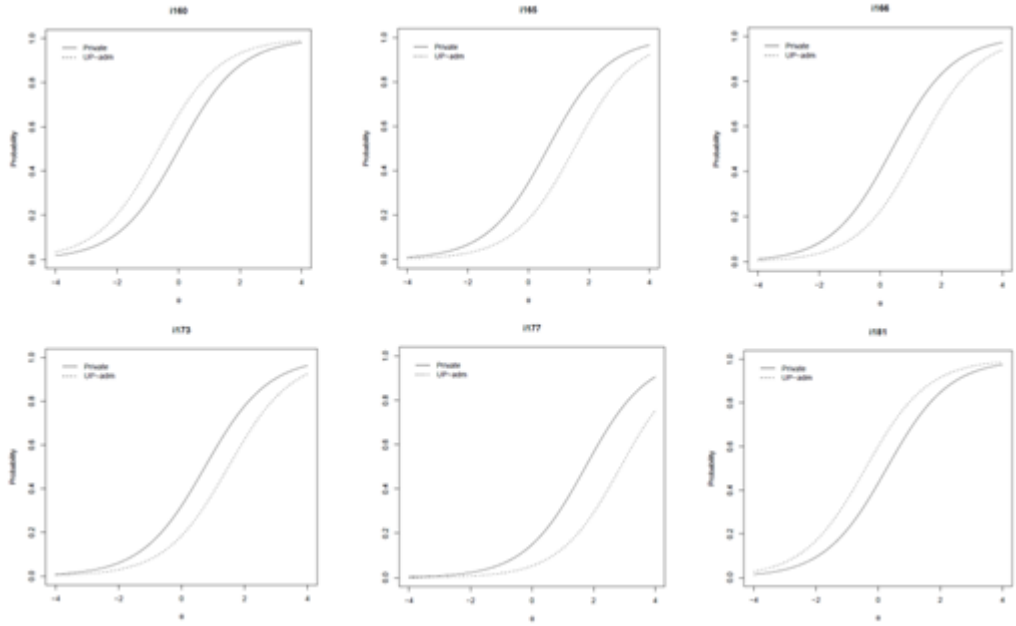

Figure 8 (d): ICCs of Items 160, 165, 166, 173, 177, and 181

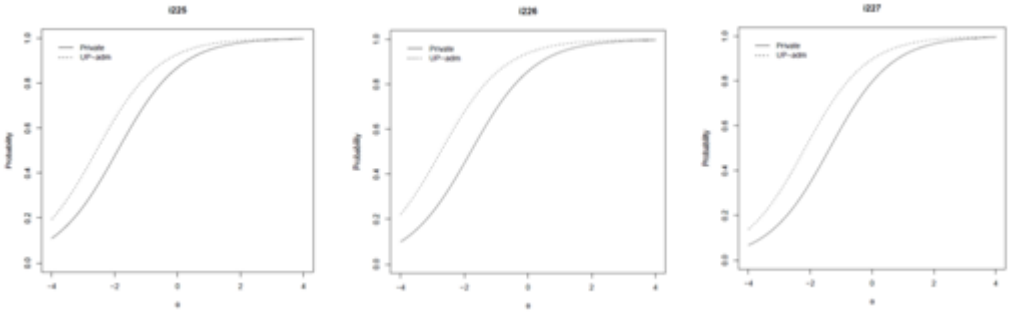

Figure 8 (e) ICCs of Items 225, 226, and 227

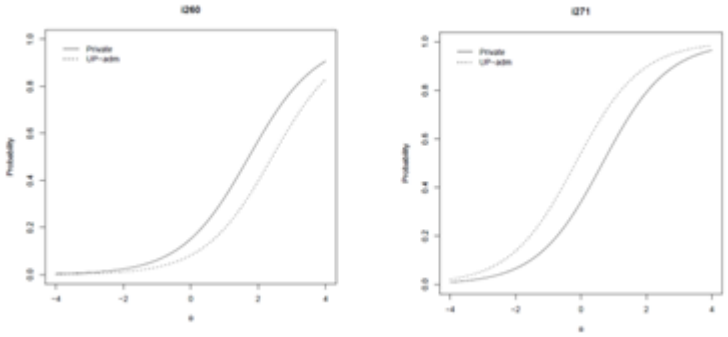

Figure 8 (f) ICCs of Items 260 and 271

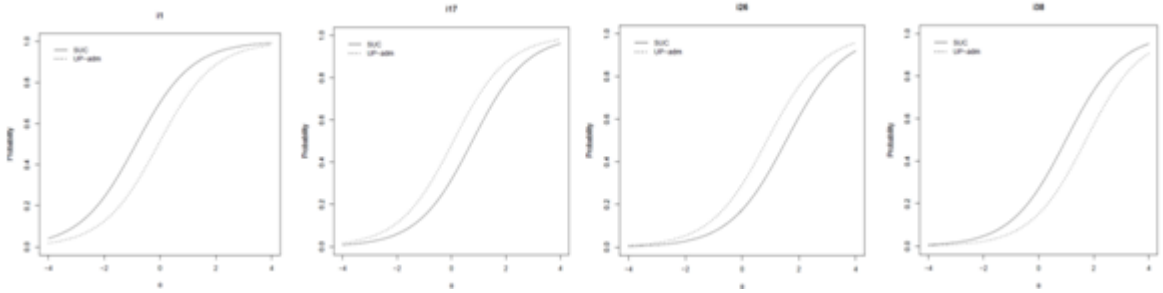

Figure 9(a): ICCs of Items 1, 17, 26, and 38

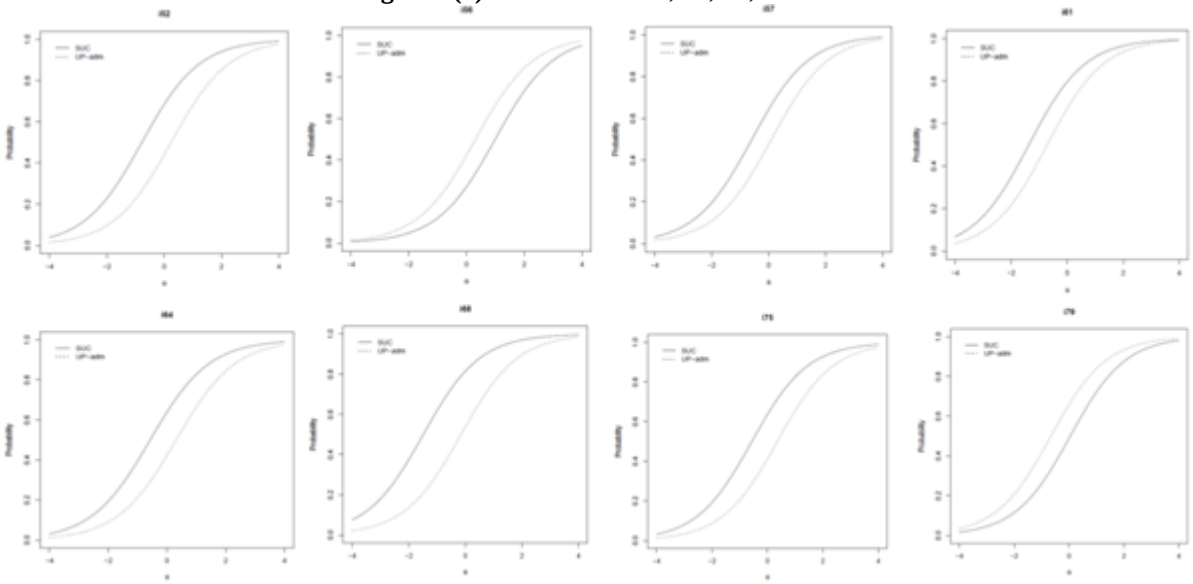

Figure 9(b): ICCs of Items 52, 56, 57, 61, 64, 68, 75, and 79

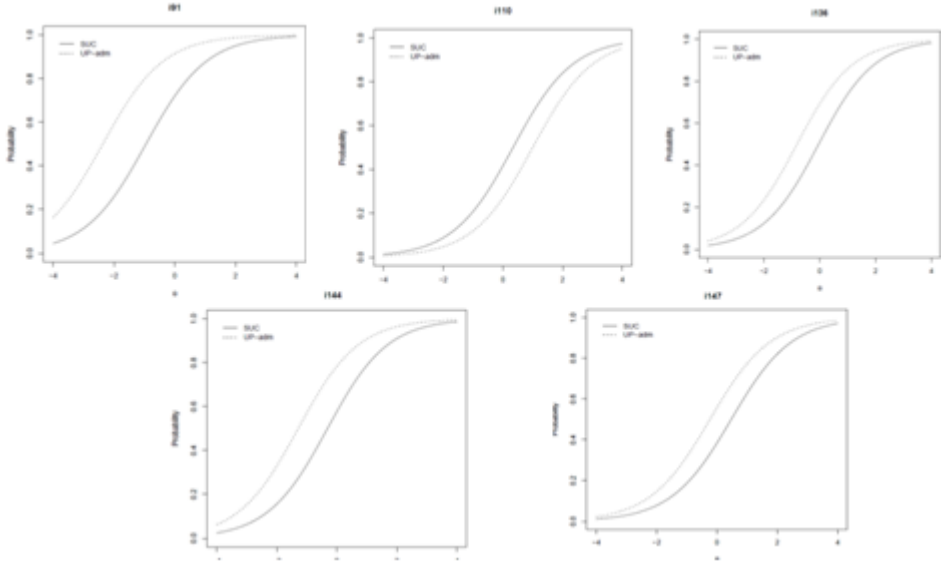

Figure 9(c): ICCs of Items 91, 110, 136, 144, and 147

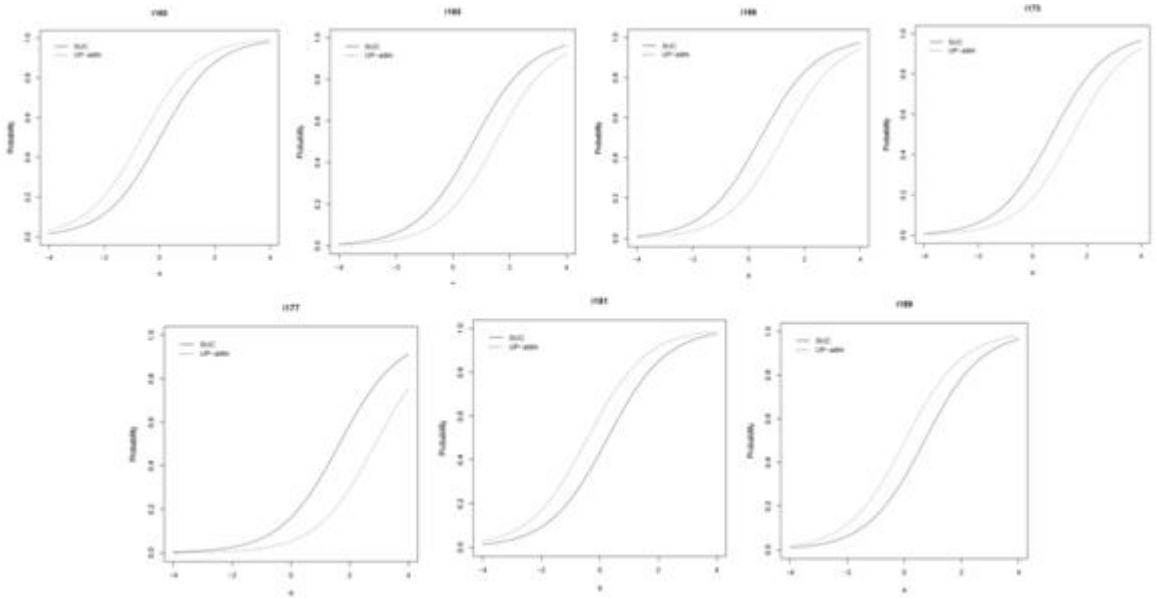

Figure 9(d): ICCs of Items 160, 165, 166, 173, 177, 181, and 189

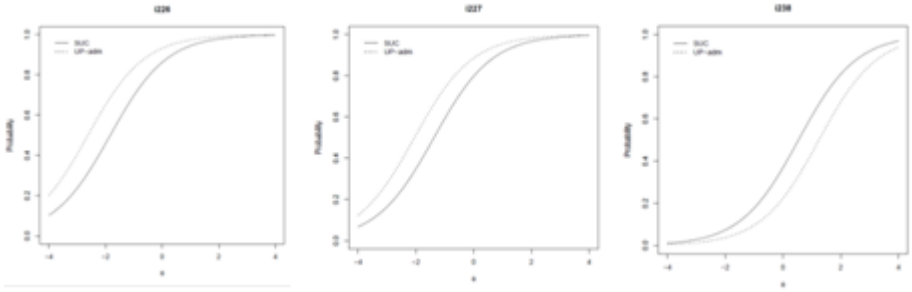

Figure 9(e): ICCs of Items 226, 227, and 238

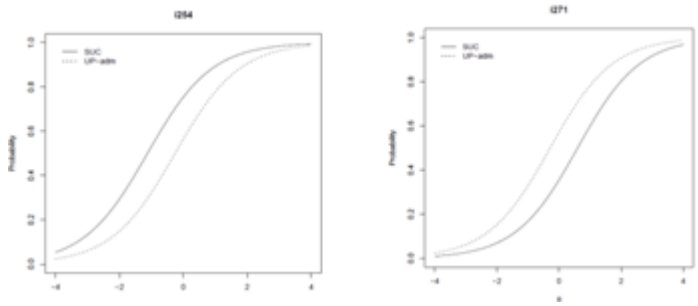

Figure 9(f): ICCs of Items 254 and 271