A Study of Gender- and Academic Major-Based Differential Item Functioning (DIF) in KEPT 2006, Mexico

Jeffrey Durand and Siwon Park

Abstract

Student grades and administrative decisions are based on a variety of assessments, including classroom and institutional examinations. An exam that is comparatively easy or difficult to a particular group of test takers can affect students' motivation and sense of self-worth. In addition, in the interests of accurate decision making and providing equal educational opportunities, these examinations should not favor one community of students over another. Towards these ends, a project was undertaken to investigate possible bias, specifically differential item functioning (DIF), in the Year 2006, Kanda English Proficiency Test (KEPT), Form Mexico. Various communities of students within the university were identified, with special attention given to the groups created by the university system and reinforced by student sociocultural affiliation. Student groups include department as well as gender. Three different DIF techniques were implemented in order to detect questions that were comparatively too easy or too difficult for a particular group. These items were then analyzed qualitatively to find out whether bias was shown.

Introduction

Student grades and administrative decisions are based on a variety of assessments, including classroom and institutional examinations. An exam that is comparatively easy or difficult to a particular group of test takers can affect students' motivation and sense of self-worth. In addition, in the interests of accurate decision making and providing equal educational opportunities, these examinations should not favor one community of students over another. Towards these ends, a project was undertaken to investigate possible bias, specifically differential item functioning (DIF), in the Year 2006, Kanda English Proficiency Test (KEPT), Form Mexico. The current paper is a research report of the project.

We begin this report by giving a brief definition of DIF. A review is followed on prior DIF studies in language testing. In the methodology section, we introduce the structure of the test data and present mathematical details of the DIF methods that we chose for the analyses. Three DIF techniques are implemented for the analyses – SIBTEST, the Mantel-Haenszel Chi Square Test, and BILOG-MG. In the section of the results, we present items that were identified with DIF through three DIF methods. The result of a cross validation analysis across different methods will be discussed at the end of the section. Finally, based on the findings, we recommend ways to improve the quality of the test by reducing the possible DIF presence with the test items.

1. Differential Item Functioning (DIF)

Item/test bias is a concept that is defined in terms of examinee groups. If all the test-takers experience a problem with a test, the test may be considered *invalid.* Yet, if the problem occurs only with a sub-group as in white vs. black

and male vs. female, we would say the test is biased. More formally, bias exists in regard to construct validity when a test is shown to measure a different psychological construct for one group over another or to measure the same trait but with differing degrees of accuracy (Reynolds, 1982). Test bias is conceptualized as individually-biased items acting in concert or in a bundle (e.g., a reading/listening text for a set of accompanying comprehension items) through a test scoring method (Shealy and Stout, 1993). Test bias is most often used in the study of racial and ethnic differences and gender differences. In recent testing literature, however, it is also common to practice bias analyses with examinees with different academic backgrounds, e.g., different college majors and different language proficiency groups. The term, differential item functioning (DIF) is currently more favored than test bias, although they are not synonymous. The decision regarding bias is made only based on the logical analysis as to why certain items are relatively more difficult or easier than others. Only based on such analyses, items will be identified as biased and will presumably be eliminated (Camilli, 1994).

The question of test bias in construct validity is of substantial concern (Messick, 1995; Reynolds, 1982). As a statistical finding, DIF signals multidimensionality with the item(s) in focus. A large DIF value suggests that on comparable examinees, the item(s) is measuring additional constructs that function differently from one group to another (Angoff, 1993; Camilli and Shepard, 1994). Depending on whether the additional construct(s) was intended to be measured, a validity account can be generated as to if the inferences of the examinee ability based on the scores are accurate and fair.

2. DIF in Language Testing

There have been studies that looked into test bias or differential item

functioning in language testing (Zumbo, 1999; Chen & Henning, 1985; Elder, 1996; Kim, 2001; Ryan & Bachman, 1992; Sasaki, 1991; and Pae, 2004). Chen and Henning (1985) attempted one of the first DIF studies in language testing. They employed a Rasch based regression procedure, in order to detect DIF in an ESL placement test for different L1 groups. Sasaki (1991) using the same method as Chen and Henning's and an additional Scheuneman's chi-square method, conducted a DIF study with the UCLA English as a Second Language Placement Examination across two distinct language learner groups of Chinese and Spanish native speakers. She found that vocabulary items with English-Spanish cognates flagged bias against the Chinese group, while test items with idiomatic expression were in favor of the Chinese group. Using the Mantel-Haenszel (MH) procedure, Ryan and Bachman (1992) examined DIF in the TOEFL and the First Certificate of English (FCE) tests across two L1 groups – Indo-European and Non-Indo-European. DIF was present with the TOEFL Listening, Structure and Written expression, and Vocabulary and Reading Comprehension sections. In the FCE, some items in the listening section presented DIF. Also using the M-H procedure, Carlton and Harris (1992) conducted an ethnicity- and gender-based DIF study for the SAT. In their study, DIF was present across gender especially with some subject matters.

Unlike those previous DIF studies with selected items, Kim's (2001) DIF study was to investigate DIF with polytomously rated scores on a speaking test. She noticed that among the skills measured in the speaking test, grammar and pronunciation functioned differently. Pae (2004), seemingly the most recent DIF study, was mainly concerned with ability groups from different academic backgrounds. Using the IRT Likelihood approach which has been highly recommended in DIF literature, he examined uniform as well as non-

uniform DIF with his test data and suggested the presence of DIF only due to group memberships in different academic backgrounds.

Methodology

1. Test Data and Structure

The data for the current research project come from Year 2006 Kanda English Placement Test (KEPT), Form Mexico. As we stated earlier, in this report, we are concerned with only the written section of KEPT with selected items, i.e., Reading comprehension, Grammar, and Listening comprehension sections. Each section consists of 35 items plus four anchor items given at the end of the written section.

As for the two DIF groups, focal and reference, we investigated DIF with two broad group memberships: 1) gender between male and female, male serving as the reference group and 2) academic majors across two to five groups depending on the type of DIF technique employed. For academic majors, depending on the grouping, one of the groups served as the reference group and naturally the other(s) as the focal group. Table 1 below presents the number of data points used for the DIF analyses based on the test section and each group membership.

Table 1. Number of data points by section and group membership

Gender		**Group**			
Male	Female	English	IC	ILC	CSK
479	1666	1372	289	228	256

CSK: Chinese, Spanish and Korean
ILC: International Language and Culture
IC: International Communication

2. DIF Detection Methods

Among various DIF detection methods, we chose and used three of which seemed most popular among DIF researchers: SIBTEST, the Mantel-Haenszel Chi Square Test, and BILOG-MG. Mathematical details of each method are presented below.

SIBTEST

SIBTEST detects DIF both in individual items and in "bundles" of items. Thus, similar items can be grouped (bundled) together to determine whether DIF exists. In this way, it is possible to test whether a sub-section of a test, such as one listening or reading passage, may be biased. Similarly, specific item types, like main idea or vocabulary in context, can be bundled together and tested as well.

SIBTEST considers both unidirectional DIF, slightly different from uniform DIF, and crossing DIF. Uniform DIF has one group scoring better than other by a constant amount. Unidirectional DIF does not assume that the amount is constant, though it is consistently in the same direction. Crossing DIF is analogous to non-uniform DIF. Item difficulties for one group are relatively higher (lower) at low ability levels and relatively lower (higher) at high ability levels. At some point in the middle, they cross, i.e. are equal.

With SIBTEST, examinees are matched by total score on the test as an indicator of ability. Two separate indices, β_{UNI} and β_{CRO}, are then calculated as functions of marginal item response functions of the reference (R) and focal (F) groups.

$$\beta_{UNI} = \int_{-\infty}^{\infty} (P_R(\theta) - P_F(\theta)) f_F(\theta) d\theta$$

$$\beta_{CRO} = \int_{\theta<\theta_C} (P_R(\theta) - P_F(\theta)) f_F(\theta) d\theta + \int_{\theta>\theta_C} (P_R(\theta) - P_F(\theta)) f_F(\theta) d\theta$$

in which θ is the overall ability and θ_C is the crossing point.

For testing unidirectional DIF, β_{UNI}, which is normally distributed with a mean of 0 and a standard deviation of 1, is used. SIBTEST uses the standard normal distribution z test to assess whether $\beta_{UNI} \neq 0$. For testing crossing DIF, β_{CRO} is calculated and significance tested through randomization.

Mantel-Haenszel Chi Square Test

Mantel-Haenszel Chi Square Test also compares a reference and focal group at matched ability levels. Randomly sampled, the reference and focal groups At each ability level, k, a 2 X 2 contingency table is created with counts of correct and incorrect answers for the reference (R) and focal (F) groups.

	Score on Item for k^{th} ability level		
Group	1	0	Total
R	A_k	B_k	n_{Rk}
F	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

A_k and C_k are independent binomial random variables. The probabilities of answering correctly (incorrectly) for those in ability group k are p_{Rk} (q_{Rk}) and p_{Fk} (q_{Fk}) for the reference and focal groups respectively. For each ability level k, the following hypothesis is tested:

$$H_0 : \frac{p_{Rk}/q_{Rk}}{p_{Fk}/q_{Fk}} = 1 \quad \text{versus} \quad H_1 : \frac{p_{Rk}/q_{Rk}}{p_{Fk}/q_{Fk}} \neq 1$$

Zwick and Ercikan (58) recommend the Mantel-Haenszel chi-square statistic to test H_0 versus H_1.

$$\text{MH CHISQ} = \frac{\left(\left| \sum_k A_k - \sum_k \mathrm{E}(A_k) \right| - \frac{1}{2} \right)^2}{\sum_k Var(A_k)},$$

in which

$$E(A_k) = n_{Rk} m_{1k} / T_k \qquad \text{and} \qquad \mathrm{Var}(A_k) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}.$$

MH CHISQ has a chi square distribution with one degree of freedom when H_0 is true. Calculating this statistic and the associated p-value reveals whether there is a significant difference between groups for each item.

BILOG-MG

An IRT analysis program, BILOG-MG (Zimkowski, Muraki, Mislevy, & Bock, 1996) is specifically designed for multiple-group IRT modeling (Embretson & Reise, 2000). It allows the users to test for DIF, based on the item difficulty parameters; however, its assumption of the DIF does not extend to the item discriminating power. The program calibrates item parameters for the two groups, DIF and Non-DIF, simultaneously. Hence, it does not require item linking procedures and provides two measures for detecting DIF with the items of interest: *b* difference test and -2 log likelihood (-2ln*L*) ratio test (a brief description about each procedure will follow shortly). The users of the program are only to decide the group reference – e.g., as one group being

the reference group and the other being the focal group.

In order to identify the baseline models, the three common IRT models, the one-, two-, and three-parameter models are run with the test in focus. If the best fit model is identified with the test data, e.g., the 2 parameter logistic model (2PLM), statistics required for DIF detection will be calculated using the model. For example, if the 2PLM is found the best fitting model, the item parameters as well as the -2 log likelihood statistics are estimated using the 2PLM. The equation of the 2PLM is in (1).

$$(1) \qquad P_i(\theta) = 1/\{1+\exp[1.7a_i(\theta - b_i)]\},$$

where

b_i is the item difficulty parameter,
a_i is the item discrimination parameter,
θ is the trait level parameter, and
1.7 is a scaling factor used to transform the metric from logistic to normal.

Using the 2PLM, two stages of analyses are performed for the measures of DIF detection mentioned earlier. First, in order to examine the magnitude of the effect that the overall language skill performance difference have on item difficulty, we take a model comparison approach using the -2 log likelihood ratio statistics (Camilli & Shepard, 1994; Thissen et al., 1988, 1993). The likelihood ratio of two models, DIF and non-DIF, can be used to obtain a test statistic such as the chi-square difference test. Using the comparison of the log likelihood of the fit of the DIF and non-DIF models, it can be statistically determined whether the difference indicates significantly better fit of the DIF model given the degrees of freedom (number of additional parameters used for the DIF model).

The fits of the 2PLMs are compared between the model with item difficulties constrained to be equal across groups (i.e., non-DIF group) and the other one in which item difficulties were allowed to vary across groups (i.e., DIF group). BILOG-MG allows performing this model comparison by producing the -2 log likelihood (-2ln*L*) statistic as in (2). Using the difference of the -2 log likelihood statistic produced by the two analyses, we can judge the goodness of fit of the models to the data and check the overall magnitude of the effect due to the overall performance difference (i.e., if the two groups can be considered independent).

(2) The-2ln*L* ratio test: $x^2(M) \approx$ -2ln(LR) = $G(2) - G(1)$, when

$M = df$, $G(2)$ = -2ln*L* (Model 2) and $G(1)$ = -2ln*L* (Models 1).

For the second stage of DIF detection, the *b*-difference test is performed. BILOG-MG estimates the threshold (*b*-parameter) difference across groups after the items have been rescaled to a common scale. In addition, standard errors (*s.e.*) between two groups (*G*2-*G*1) as in (3) are generated for each difference. We can determine whether the threshold difference is statistically different from zero – i.e., no difference.

(3) $s.e._{G2-G1} = \sqrt{\text{s.e.}^2_{G2} + \text{s.e.}^2_{G1}}$

A critical ratio test can be performed with the differences between the *b*-parameters over the *s.e.* for the item of interest. This procedure which is equivalent to Muraki and Engelhard's *standardized index of bias (SIB)* test uses two standard deviations in group ability differences as the criterion to

judge whether or not DIF is present with the item (Thissen et al., 1993).

(4) critical ratio test: $b_2 - b_1 / \sqrt{\text{s.e.}^2_{G2} + \text{s.e.}^2_{G1}}$,

where

2 represents the reference group, and

1 represents the focal group.

Results

1. Gender-based DIF

SIBTEST

On the reading section, five items show evidence of DIF for gender at the p=.05 level. Items 10, and 16, as well as anchor item 110, favor males. Items 12 and 25 favor females. On the grammar section, eight items show evidence of DIF at the p=.05 level. Items 40, 48, 64, and 70, along with anchor item 106 favor males, while items 38, 62, and 63 favor females. On the listening section, five items show evidence of DIF at the p=.05 level. Items 76, 81, and 93 favor males while items 92 and 95 favor females.

MANTEL-HAENSZEL CHI SQUARE TEST

On the reading section, five items show evidence of DIF for gender at the p=.05 level. Items 10, and 16, as well as anchor item 110, favor males. Items 12 and 25 favor females. These are the same items that SIBTEST shows as having DIF. On the grammar section, eight items show evidence of DIF. Items 40, 48, 60, and 70, along with anchor item 106 favor males, while items 38, 62, and 63 favor females. These are almost the same as with SIBTEST, the only difference being item 60 instead of item 64 favoring males. On the listening

section, four items show evidence of DIF. Items 76 and 93 favor males while items 92 and 95 favor females. The only difference from SIBTEST is that item 81 no longer favors males.

Table 2. Items with statistically significant gender-based DIF

	SIBTEST		**M-H Test**		**BILOG-MG**	
In favor of:	Male	Female	Male	Female	Male	Female
Reading	10 16 110	12 25	10 16 110	12 25	*none*	5
# of DIF items	3	2	3	2	0	1
Grammar	40 48 64 70 106	38 62 63	40 48 60 70 106	38 62 63	*none*	*none*
# of DIF items	5	3	5	3	0	0
Listening	76 81 93	92 95	76 93	92 95	*none*	*none*
# of DIF items	3	2	2	2	0	0

Note: Items highlighted are DIF items that are identified only by the method in focus.

BILOG-MG

On the reading section, only Item 5 was identified with DIF favoring females. As Table 2 shows, no other item displayed DIF on the grammar and

listening sections. This result is rather surprising considering the number of DIF items flagged by the other two techniques. We will return to this issue in the section of cross validation.

2. Major-based DIF

SIBTEST

Chinese, Spanish, and Korean (CSK) majors were combined into one group as they study English together. The relative small number of each major also makes their separate testing difficult. As a result, paired combinations of four different majors were analyzed.

English versus Chinese, Spanish and Korean (CSK)

On the reading section, six items show evidence of DIF. Items 6, 31, and 35 favor English majors while items 7, 8, and 24 favor CSK majors at the .05 level. On the grammar section, seven items show evidence of DIF: items 37, 44, 54, and 60 favor English majors, while items 38, 50, and 51 favor CSK majors. On the listening section, five items show DIF. Items 81, 87, and 94 favor English majors, while items 92 and 98 favor CSK majors.

English versus International Language and Culture (ILC)

On the reading section, eleven items show evidence of DIF. Items 10, 12, 19, 25, and 35, as well as anchor item 111, favor English majors. Items 2, 22, 23, 28, and 34 favor ILC majors. On the grammar section, six items show evidence of DIF. Items 36, 37, 53, and 69, along with anchor item 107, favor English majors. Item 52 is the only one to favor ILC majors. On the listening section, three items show DIF. Item 105 favors English majors while items 87 and 90 favor ILC majors.

English versus International Communication (IC)

On the reading section, three items show DIF. Items 12 and 35 favor English majors. Item 13 favors IC majors. On the grammar section, item 41 and anchor item 107 favor English majors while items 61 and 64 favor IC majors. On the listening section, six items show DIF. Items 85 and 90 and anchor item 115 favor English majors, while items 93, 95, and 96 favor IC majors.

International Language and Culture (ILC) versus Chinese, Spanish and Korean (CSK)

On the reading section, six items show DIF. Items 2, 6, 22, and 34 favor ILC majors. Items 8 and 24 favor CSK majors. On the grammar section, four items show DIF: items 39 and 52 favor ILC majors and items 50 and 69 favor CSK majors. On the listening section, four items show DIF. Items 87, 94, and 100 favor ILC students while item 105 favors CSK students.

International Communication (IC) versus Chinese, Spanish and Korean (CSK)

On the reading section, four items show DIF. Item 6 favors IC students while items 12, 21, and 24 favor CSK students. On the grammar section, seven items show DIF. Items 37, 43, 44, 60, 61, and 64 favor IC students while item 51 favors CSK students. On the listening section, five items show DIF. Items 81, 94, 95, and 96 favor IC students while item 92 favors CSK students.

International Communication (IC) versus International Language and Culture (ILC)

On the reading section, two items, 22 and 28, favor ILC students. No items favor IC students. On the grammar section, seven items show DIF. Items 37,

57, 61, 64, and 69 favor IC students. Items 49 and 51 favor ILC students. And on the listening section, four items show DIF. Items 103 and 105 favor IC students. Items 87 and 90 favor ILC students.

MANTEL-HAENSZEL CHI SQUARE TEST

English versus Chinese, Spanish and Korean (CSK)

On the reading section, nine items show evidence of DIF, three additional items to SIBTEST. Items 4, 6, 10, 31, and 35 favor English majors while items 7, 8, 21, and 24 favor CSK majors at the .05 level. On the grammar section, nine items show evidence of DIF. Items 36, 37, 39, 44, 54, 60, and 61 favor English majors, while items 50 and 51 favor CSK majors. On the listening section, seven items show DIF. Items 81, 85, 89, 94, and 100, along with anchor item 117, favor English majors, while item 98 favors CSK majors.

English versus International Language and Culture (ILC)

On the reading section, eight items show evidence of DIF. Items 10, 14, and 25, as well as anchor item 111, favor English majors. Items 2, 22, and 34, along with anchor item 113, favor ILC majors. On the grammar section, nine items show evidence of DIF. Items 36, 37, 47, 53, 57, 60, and 69, along with anchor item 107, favor English majors. Item 52 is again the only one to favor ILC majors. On the listening section, four items show DIF. Items 103 and 105, along with anchor item 114, favor English majors, while item 71 favors ILC majors.

English versus International Communication (IC)

On the reading section, three items show DIF. Items 12 and 35 favor English

majors. Item 13 favors IC majors. These are exactly the same results as with SIBTEST. On the grammar section, item 41 and anchor item 107 favor English majors, while item 64 favors IC majors. On the listening section, six items show DIF. Items 85 and 90 and anchor item 115 favor English majors, while items 93, 95, and 96 favor IC majors. These also are the same as SIBTEST.

International Language and Culture (ILC) versus Chinese, Spanish and Korean (CSK)

On the reading section, five items show DIF. Items 2, 6, and 22 favor ILC majors. Items 8 and 24 favor CSK majors. On the grammar section, five items show DIF: items 39, 61, and 52 favor ILC majors and items 50 and 69 favor CSK majors. On the listening section, only two items show DIF. Item 94 favors ILC students while item 105 favors CSK students.

International Communication (IC) versus Chinese, Spanish and Korean (CSK)

On the reading section, six items show DIF. Item 6, 9, 15, and 17 favor IC students while items 12 and 21 favor CSK students. On the grammar section, nine items show DIF. Items 37, 54, 57, 60, 61, and 64 favor IC students while items 50 and 51 favor CSK students. On the listening section, seven items show DIF. Items 81, 94, 95, and 96, as well as anchor items 114 and 117, favor IC students while item 92 favors CSK students.

International Communication (IC) versus International Language and Culture (ILC)

On the reading section, three show DIF. Item 25 and anchor item 111 favor IC students. Item 28 favors ILC students. On the grammar section, four

items show DIF. Items 37, 57, 64, and 69 all favor IC students. Finally, on the listening section, seven items show DIF. Items 76, 93, 96, and 105 favor IC students. Items 87 and 90 favor ILC students.

BILOG-MG

Table 3.2 notes the number of major-based DIF items identified by BILOG-MG. Only three items are DIF present – Item 112 on the reading section, Items 105 and 117 from the listening section. Item 112 repeatedly occurs as a DIF item with groups of English vs. CSK, English vs. ILC, and English vs. IC consistently favoring English over the other groups. Surprisingly, no items are identified with DIF on the grammar section. On the listening section, Item 105 shows DIF, favoring English over ILC; Item 117 favoring CSK over ILC and IC over ILC.

Table 3.1 Items with statistically significant major-based DIF by SIBTEST and M-H Test

	SIBTEST												M-H Test											
comparison	E vs. CSK		E vs. ILC		E vs. IC		ILC vs. CSK		IC vs. CSK		IC vs. ILC		E vs. CSK		E vs. ILC		E vs. IC		ILC vs. CSK		IC vs. CSK		IC vs. ILC	
In favor of:	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
Reading	6			2	12		2		6			22	4			2	12		2		6		25	
		7	10			13	6			12		28	6		10			13	6		9			28
		8	12		35			8		21				7	14		35			8		12	111	
		24	19				22			24				8					22		15			
	31			22				24					10			22				24	17			
	35			23			34							21	25							21		
			25											24		34								
				28											111									
				34									31			113								
			35										35											
			111																					
# of DIF items	3	3	6	5	2	1	4	2	1	3	0	2	5	4	4	4	2	1	3	2	4	2	2	1
Grammar	37		36		41		39		37		37		36		36		41		39		37		37	
		38	37			61		50	43			49	37		37			64		50		50		
	44			52		64	52		44			51			47		107		52			51		
		50	53		107			69		51	57		39			52			61		54		57	
		51	69						60		61		44		53					69	57		64	
	54		107						61		64			50	57						60		69	
	60								64		69			51	60						61			
													54		69						64			
													60		107									
													61											
# of DIF items	4	3	5	1	2	2	2	2	6	1	5	2	7	2	8	1	2	1	3	2	6	2	4	0
Listening	81			87	85		87					87	81			71	85		94		81		76	
	87			90	90		94					90	85				90				84			
		92	105			93	100				103		87		103			93		105		92		87
	94					95		105			105		89		105			95			95		93	90
		98				96							94		114			96			96		96	
					115									98			115				114			
													100								117		105	
													107											
# of DIF items	3	2	1	2	3	3	3	1	0	0	2	2	7	1	3	1	3	3	1	1	6	1	4	2

E: English; CSK: Chinese, Spanish and Korean; ILC: International Language and Culture; IC: International Communication

In favor of: 1 - first group and 2 - second group

Table 3.2 Items with statistically significant major-based DIF by BILOG-MG

	BILOG-MG											
comparison	E vs. CSK		E vs. ILC		E vs. IC		ILC vs. CSK		IC vs. CSK		IC vs. ILC	
In favor of:	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
Reading	112	none	112	none	112	none	none	none	none	none	None	none
# of DIF items	1	0	1	0	1	0	0	0	0	0	0	0
Grammar	none	none	none	none	none	none	none	none	none	none	None	none
# of DIF items	0	0	0	0	0	0	0	0	0	0	0	0
Listening	none	none	105	none	none	none	none	117	none	none	117	none
# of DIF items	0	0	1	0	0	0	0	1	0	0	1	0

In favor of: 1 - first group and 2 - second group

3. Cross Validation

SIBTEST vs. the Mantel-Hanszel Chi Square Test vs. BILOG-MG

SIBTEST and the Mantel-Hanszel chi square test, which calculate DIF in similar ways, produced similar results. For gender-based DIF, they both identified 16 of the same items as showing DIF. The two techniques had only three differences. SIBTEST showed that listening item 81 and grammar item 64 favor males, and the Mantel-Hanszel chi square test showed that grammar item 60 favors males as well. For two of these three items, grammar items 60 and 64, the p-value for the technique not showing DIF was still less than .10.

For major-based DIF, the picture is much the same, except for *English versus CSK students*. Here, both methods identified 15 of the same items as showing DIF. Another 13 items were detected by one method but not the other. For only 3 of these 13 items (23 percent), the p-value for the technique not showing DIF was less than .10, but for 8 items, the difference was large, more than .20. For *English versus ILC* students, 13 items were detected by

both methods and another 15 items by only one method. For 9 of these 15 (60 percent), the p-value for the technique not showing DIF was still less than .10. For *English versus IC* majors, 12 items were identified by both techniques. One additional item, grammar item 61, was identified by SIBTEST only, but the difference in p-values is .026. For *ILC versus CSK* students, 11 items were identified by both methods, and another 4 by one method only. For 3 of these 4 items (75 percent), the p-value for the technique not showing DIF was still less than .10. For *IC versus CSK* students, 13 items were identified by both methods. Another 12 items were identified by one method only, but of these, 7 items (58 percent) had p-values of less than .10 by the other technique. Finally, for *IC versus ILC* students, 8 items were identified by both methods and an additional 11 items by one method only. Of these 11 items, 7 items (64 percent) had p-values from the other technique of less than .10.

BILOG-MG did not identify as many DIF items as SIBTEST and the M-H test did. This finding certainly deserves further attention to understand the possible causes. One explanation of such disparity in DIF detection sensitivity across different techniques/programs may be due to the mathematical adjustment for Type I error that was differently implemented within each program. For instance, researchers recommend using the Bonferroni adjustment for the judgment of DIF presence. For instance, when SIBTEST is used for DIF detection, it may be desirable to adjust the criterion level of significance (i.e., 0.05) by dividing it by the number of items to correct the possible Type I Error. In the current study, we did not apply such adjustment technique to our DIF investigation and that may be responsible for a large number of DIF items detected by SIBTEST in contrast to BILOG-MG. In our subsequent reports on DIF with KEPT, we detail this issue by exploring

other possibilities responsible for the different DIF findings across different programs/techniques (Durand & Park, 2007; Park, Durand, & Batty, 2006).

Conclusion

The current study examined DIF with the written section of the 2006 Kanda English Proficiency Test (KEPT) using three DIF programs. SIBTEST and the Mantel-Hanszel Chi Square Test identified rather a large number of DIF items across three sections of reading, grammar, and listening. Unlike the other two techniques, BILOG-MG was resulted with a small number of DIF items. As stated earlier, this disparity deserves further consideration so that more in depth understanding of each technique can be facilitated and more precise interpretation of the DIF findings can be produced.

As we discussed earlier, DIF does not necessary mean bias. One can argue for the presence of bias with a DIF item only through a qualitative verification of such bias with the content or skill of the item that is elicited. The DIF items detected throughout this study also must go through such endeavor so that such items can be revised properly or disregarded entirely.

References

Angoff, W.H. (1993). Perspectives on Differential Item Functioning Methodology. In P. W. and H. Wainer (eds), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* (Vol. 4) Thousand Oaks, CA: Sage Publications.

Carlton, S.T., & Harris, A.M. (1992). Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: gender and majority /minority group comparisons. *ETS Research Report*, 92–64. Princeton, NJ: ETS.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–63.

Durand, J., & Park, S. (forthcoming). A report of Differential Item Functioning in item bundles: A Research Report.

Elder, C. (1996) The effect of language background on foreign language test performance: the case of Chinese, Italian, and modern Greek. *Language Learning, 46*, 233–82.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum.

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

Kim, M. (2001). Detecting DIF across different language groups in a speaking test. *Language Testing, 18*(1), 89–114.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

McNamara, T., O'Loughlin, K., & Wigglesworth, G. (1993). *Bias analysis in Rasch programs: analyzing interactions between judges, items and persons in performance assessment settings.* Paper presented at the annual conference of the American Association for Applied Linguistics, Atlanta GA, April.

Park, S., Durand, J., & Batty, A. (2006). A Research Report on DIF for Kanda English Proficiency Test, 2006: Analyses using Simultaneous Bias Test (SIBTEST), the Mantel-Haenszel Chi Square Test, and BILOG-MG.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178-208). New York: John Wiley.

Ryan, K., & Bachman, L.F. (1992). Differential Item Functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12–29.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*(2), 95–111.

Shealy, R.T., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–94.

Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–69). Hillsdale, NJ: Lawrence Erlbaum.

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

Zimkowski, M. E., Muraki, E., Mislevy, R., J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

Zwick, R., & Ericikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. *Journal of Educational Measurement, 26*(1), 55-66.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.