Detecting DIF across Different Language and Gender Groups in the MELAB Essay Test using the Logistic Regression Method

Taejoon Park
Teachers College, Columbia University

This study has investigated differential item functioning (DIF) on the writing subtest of the Michigan English Language Assessment Battery (MELAB). A total of ten writing prompts were examined using a three-step logistic regression procedure for ordinal items. An English Language Ability (ELA) variable was created by summing the standardized MELAB Grammar, Cloze, Vocabulary, Reading (GCVR), and Listening scale scores. This ELA variable was used to match examinees of different language (Indo-European vs. non–Indo-European) and gender (male vs. female) groups. The results of this study showed that although a few prompts were initially flagged because of statistically significant group effects, the effect sizes were far too small for any of those flagged prompts to be classified as having an important group effect.

That a test should not be biased is an important consideration in the selection and use of any psychological test. That is, it is essential that a test is fair to all applicants, and is not biased against a subgroup of the test-taker population. Bias can result in systematic errors that distort the inferences made from test scores. Test bias occurs when items contain sources of difficulty that are irrelevant or extraneous to the construct or ability being measured, and these extraneous or irrelevant factors affect performance (Zumbo, 1999). This issue of test bias has been the subject of a great deal of recent research in the field of educational measurement, and a statistical procedure called Differential Item Functioning (DIF) analysis has become an indispensable part of test development and validation. In the context of the impromptu essay test, when a prompt or topic is biased against a certain subgroup(s) of examinees, it could potentially distort the meaning of the essay score for different examinee subgroups and become a potential threat to the validity of score interpretations (Sheppard, 1982). For this reason, DIF screening must become an indispensable part of test development and validation.

MELAB Essay Test

Direct assessments of writing ability such as the MELAB essay test have considerable appeal because, unlike their less direct, multiple-choice counterparts, they actually require examinees to write, not merely to recognize the conventions of standard written English. Along with their appeal, however, these tests carry a special burden that does not encumber traditional multiple-choice tests. That is, they typically consist of a single writing task, which limits the generalizability of the results and may disadvantage test takers who happen to have little interest or background in the assigned topics (Weigle, 2000). Thus, it is incumbent upon test developers to devise essay prompts that are fair to all examinees and as comparable to one another as possible. The aim is that no test takers will be unfairly disadvantaged by being

administered a prompt whose content is so unfamiliar or unengaging as to hinder the demonstration of their writing ability.

In the MELAB essay test, which is the first part of the battery, examinees are given 30 minutes to write on a single topic they choose from a set of two topics. At any given time approximately 25 sets of topics (or 50 individual topics) are being used as MELAB prompts. Since only one essay prompt is administered per examinee in the composition section, it becomes a very important issue to ensure that each prompt is as fair as possible to any subgroups of examinees who take the same writing prompt from the pool of prompts. The MELAB program already does much to ensure the fairness of the writing prompts. According to the MELAB technical manual (English Language Institute, 2003, pp. 33–34):

> Topic writers are advised to consider various constraints of the test situation, expectations about the writing outcome, and various characteristics of the examinees when formulating possible MELAB writing topics. Because it is expected that the text produced by the examinee will be at least 150 words long, prompt writers are advised to develop topics that are broad enough that someone can write at length. Certain topics are avoided, specifically those that might elicit formulaic or previously prepared responses . . . . Characteristics of the examinees that are considered when devising topics are that the examinees come from a range of linguistic and cultural backgrounds, have various educational backgrounds, and are different ages. Topics are developed to be accessible and attractive to a range of young adult and adult examinees. Topics are avoided that could be considered politically or culturally objectionable or limited, or that require the examinee to draw on specialized knowledge of a culture, a field, or a discipline. Topics may call upon the personal experience, attitudes, or general knowledge of the examinees.

Differential Item Functioning (DIF)

DIF addresses the issue of differential validity across groups by examining the performance of two groups of interest after the groups have been matched on some criterion (Dorans & Holland, 1993). According to O'Neill and McPeek (1993), "The fundamental principle of DIF is simple: Examinees who know the same amount about a topic should perform equally well on an item testing that topic regardless of their sex, race, or ethnicity" (p. 256).

DIF occurs when examinees of equal ability, but with different group membership, have unequal probabilities of success on an item (Angoff, 1993; Clauser & Mazor, 1998). As described above, in order to minimize the likelihood of this situation, test developers attempt to craft prompts that are as nearly equivalent as possible and thus, to the extent possible, ensure all essay prompts function similarly for all test takers. In this way, any between-group difference in performance on a prompt will be due to construct-relevant factors rather than to influences that are irrelevant to the assessment of writing ability. In addition, though, it is desirable to use statistical techniques (i.e., DIF procedures) for investigating potential bias, especially in high-stakes testing situations.

The identification of a satisfactory DIF procedure for essay prompts is not an easy undertaking for a number of reasons. For multiple-choice items, well-established methods exist for detecting items that are differentially difficult for certain subgroups of test takers. However, there are currently no entirely satisfactory, well-researched comparable procedures

for determining whether essay prompts are differentially difficult for matched subgroups of test takers.

As pointed out by Lee, Breland, and Muraki (2002), one of the most important challenges in conducting DIF on essay tests is to find an appropriate variable to use for matching examinees of two different language groups on their writing ability. In the DIF procedure, this overall matching must be accomplished before between-subgroup performance comparisons can be made for individual items. For standardized multiple-choice measures, the total score on the test typically serves this function. For direct writing assessments, a comparable internal matching criterion is not usually available because most high-stakes writing tests such as the MELAB essay test consist of only a single writing prompt. In such cases, the usual strategy is to use an external matching criterion such as scores on multiple-choice tests that measure similar knowledge, skills, or abilities (Lee, Breland, & Muraki, 2002).

In the present study, an external matching variable was created by summing the standardized scale scores from the two multiple-choice subtests of the MELAB (Grammar, Cloze, Vocabulary, Reading [GCVR], and Listening) based on a recommendation by Penfield and Lam (2000). The underlying assumption is that if examinees have high ELA measured by two parts of the test combined, they should perform well overall on the essays, and vice versa (more detailed information on the matching variable is provided later in the Method section).

Uniform and Nonuniform DIF

An item shows uniform DIF when the performance of one group is always superior to another group for each ability level. All DIF procedures are capable of detecting uniform DIF. An item shows nonuniform DIF when the performance of one group is dependent upon ability level. Thus, nonuniform DIF displays an interaction between ability level and group membership. Because of this interaction, nonuniform DIF is much more difficult to interpret.

The identification of nonuniform DIF in polytomous items may be more important than the identification of nonuniform DIF in dichotomous items (Spray & Miller, 1994). There are many possible group-by-response-by-score interactions that can manifest in the polytomous case. For a polytomous DIF detection method to be useful, the power for detecting such interactions should be sufficiently large. Many DIF procedures are not capable of detecting nonuniform DIF, and thus one criterion for judging the usefulness of DIF procedures is its ability to detect nonuniform DIF.

DIF Detection Methods for Polytomous Items

Although there are several statistical methods available for the detection of differentially functioning items that are scored dichotomously (e.g., IRT approaches, the Mantel-Haenszel statistic, and SIBTEST), these methods are not directly applicable to polytomously scored items. Recently, various methods of investigating DIF in polytomous items have been developed, including logistic regression (Zumbo, 1999), logistic discriminant function analysis (Miller & Spray, 1993), Polytomous IRT (Muraki, 1999), the Generalized Mantel-Haenszel procedure (Zwick, Donoghue, & Grima, 1993), the polytomous SIBTEST procedure (Chang, Mazzeo, & Roussos, 1995), and the standardization method (Dorans & Schmitt, 1991). Among these methods, however, those requiring an internal criterion (e.g., polytomous IRT and polytomous SIBTEST) are not feasible for this study. Moreover, methods such as the Generalized Mantel-Haenszel procedure and the standardization method

lack the power to detect nonuniform DIF (Miller & Spray, 1993), which may be even more important when dealing with polytomous items due to the multiple ways in which item scores interact with the total score (Spray & Miller, 1994).

For the present study, the logistic regression DIF method was selected because of its ability: (1) to detect both uniform and nonuniform DIF, and (2) to supplement the statistical test with means by which the practical significance of DIF can be examined. The logistic regression method employed in this study is described in detail in the Method section.

The primary purpose of this study is to investigate DIF across different language and gender groups on the MELAB essay test, using the logistic regression method.

Method

Sample

Included in the original sample of data available for this study were 5991 examinees who responded to 75 different topics. Sixty-five prompts with small sample sizes ($n < 140$) were dropped from the analysis. Of the 2,269 examinees used for the logistic regression analyses, 686 were males and 1583 were females. A total of 575 examinees were categorized as an Indo-European language group and 1694 as a Non-Indo-European language group. The most frequently reported first language is Tagalog/Filipino (27.90%), followed by Cantonese/Mandarin (13.84%), Korean (10.31%), Farsi/Persian (4.72%), Malayalam (4.58%), Russian (4.23%), Spanish (2.69%), Arabic (2.42%), Urdu (2.12%), Japanese (2.03%), English (1.85%), Somali (1.85%), Hindi (1.63%), Tamil (1.41%), Other Asian (1.37%), Romanian (1.32%), Vietnamese (1.28%), Punjabi (1.15%), Indonesian (1,06%), Other-African (0.93%), Bengali (0.88%), Gujarati (0.88%), Amharic (0.71%), French (0.71%), Polish (0.62%), Turkish (0.53%), Thai (0.48%), Portuguese (0.44%), Albanian (0.40%), Bulgarian (0.35%), German (0.31%), Tibetan (0.31%), Ukrainian (0.31%), Croatian (0.22%), Ormo (0.22%), Serbian (0.22%), etc.

Instruments

Data analyzed included scores on the Writing, Grammar, Cloze, Vocabulary, Reading (GCVR), and Listening subtests of the MELAB. The MELAB Writing score is based on two independent readings and holistic ratings of the essay response on a ten-step scale (53, 57, 63, 67, 73, 77, 83, 87, 93, or 97). The Writing test score is basically the average of the two reader ratings, and a third reader is also used when the two initial ratings differ by more than one scale point (See Appendix A for descriptions of each score level). The Listening and GCVR subtests have a score range from 30 to 100 and from 15 to 100, respectively.

A matching variable, named ELA score, was created by (a) taking all the examinees who took the same writing prompt; (b) standardizing the scale score of the Listening, and GCVR subtests separately based on the total examinee samples for a specific prompt; and (c) summing the standardized scores of the two subtests for each examinee. One might argue that the GCVR score alone (i.e., without the Listening scores combined) could be a more valid matching variable for the writing ability. However, when the scale scores from the two subtests (i.e., GCVR and Listening) were standardized and combined, the correlation between the essay score and the matching criterion was maximized. Thus, a decision was made to create a matching variable by summing the standardized scale scores from the two subtests for each of the prompts analyzed in this study.

Computer Programs

Stata version 8.0 was used to conduct all the statistical analyses (e.g., descriptive statistics, independent samples t-tests, and ordinal logistic regression) used in this study. The Microsoft Excel database software program was used for the purpose of data management.

Data Analyses

T-tests

Independent samples t-tests were computed using examinees' essay raw scores as the independent variable and language backgrounds (Indo-European vs. non-Indo-European) and gender (male vs. female) as the independent variables. The assumption of homogeneous variance was determined through use of the Levene statistic. If the Levene statistic was significant, the t-test result assuming unequal variance was interpreted; if the Levene statistic was not significant, the t-test result assuming equal variance was interpreted. For all independent samples t-tests, an $\alpha = 0.05$ was used and two-tailed results were examined. Practical significance of the results was determined by computing Cohen's *d* (standardized mean difference effect size) and using the cutoffs of 0.2, 0.5. 0.8 as small, medium and large, respectively (Cohen, 1992).

Logistic Regression for DIF

French and Miller (1996) and Zumbo (1999) demonstrated that the logistic regression procedure (Hosmer & Lemeshow, 1989) could be extended to detect DIF in polytomous items. As pointed out by Lee, Breland, and Muraki (2002), logistic regression has two main advantages over linear regression. The first is that the dependent variable does not have to be continuous, unbounded, and measured on an interval or ratio scale. In the case of the MELAB essay test data, the dependent variable (the essay score) is discrete and bounded between 53 and 97. The second advantage is that it does not require a linear relationship between the dependent and independent variables. Thus, logistic regression allows for the investigation of the group membership effect on the dependent variable, whether the relationship between the dependent and the independent variables is linear or nonlinear. When a dependent variable is discrete and bounded while the independent variable is continuous, a nonlinear relation is likely to exist among the variables (Lee, Breland, & Muraki, 2005). For these reasons, the logistic regression method is most appropriate for the study.

In the case of the MELAB essay test, each examinee's essay is scored on an ordinal scale, and thus the ordinal logistic regression was used in this study. The ordinal logistic regression estimates the cumulative probability and describes the relationships between each variable of the model. For the detection of DIF, the full ordinal logistic regression model used in this study is as follows:

$$\text{logit}\,[P(Y \le k)] = \alpha_k + b_1(\text{ELA}) + b_2(\text{Group}) + b_3(\text{ELA*Group})$$

where Y = the natural log of the odds ratio,
k = 0, 1, 2 . . . *m*, where *m* is the number of categories in the ordinal scale,
ELA is the matching variable used in this study, and
ELA*Group is the matching variable by group membership interaction variable.

More specifically, a three-step modeling process based on ordinal logistic regression (Zumbo, 1999) was used as the main method of analysis. That is, the ordinal logistic

regression analysis was conducted in the following three steps: step 1, only the matching variable or the conditioning variable (i.e., ELA scores) was entered into the regression equation, as in, logit [P (Y ≤ k)] = $\alpha_k + b_1$(ELA); step 2, the group membership variable was entered into the regression equation, as in, logit [P (Y ≤ k)] = $\alpha_k + b_1$(ELA) + b_2(Group); and step 3 (i.e., the full model), the interaction term (i.e., English Language Ability-by-Group) was finally added to the regression equation, as in, logit [P (Y ≤ k)] = $\alpha_k + b_1$(ELA) + b_2(Group) + b_3 (ELA*Group).

Using the three-step modeling process described above, the logistic regression method compares the fit of the augmented model (entering additional variables hierarchically into the model) to that of the compact model. If the first augmented model including the "Group" variable fits the data, this suggests that the prompt shows DIF due to group membership. That is, if the null hypothesis of "$b_2 = 0$" is rejected and "$b_3 = 0$" is tenable, the prompt shows uniform DIF. Likewise, if the fully augmented model including the "ELA*Group" interaction variable fits the data, this indicates that both ELA and group membership contribute to DIF. In other words, the null hypothesis of "$b_3 = 0$" is rejected, and nonuniform DIF is present in the prompt.

In order to gauge the amount of the group difference (if any), *p*-values for the Chi-square test were used along with R^2 effect size estimates, which provide information about the practical significance of DIF (Zumbo, 1999).

In the present study, the uniform R^2 effect size is basically an increased portion of R^2 after entering the dummy language group variable into the ELA-only regression model (i.e., step-1 model). The nonuniform effect size is an increased portion of R^2 after adding the interaction term in the step-2 model. The total effect size is the aggregate of the uniform and nonuniform effects. There seems to be a lack of agreement over just what constitutes small or negligible, moderate or medium, or large effects. Cohen (1988) considered R^2 effect sizes of 0.02, 0.13, and 0.26 as "small," "medium," and "large" effect sizes, respectively, which can also be linked to the standardized group mean differences (i.e., Cohen's *d*) of 0.20, 0.50, and 0.80 in standard deviation units. Zumbo (1999) suggested that, for an item to be classified as displaying DIF (i.e., an aggregate of uniform and nonuniform DIF), the 2-degree of freedom Chi-square test between steps 1 and 3 have to have a *p*-value less than or equal to 0.01 and the R^2 difference between them should be at least 0.13. Zumbo's classification scheme of the R^2 values of 0.13 corresponds to a "medium" R^2 effect size in Cohen's standard. Jodoin and Gierl (2001) suggested that R^2 differences of 0.035, 0.035 to 0.070, and greater than 0.070 be considered as "negligible," "moderate," and "large" effects.

Results

Descriptive Statistics

Descriptive statistics of the ten prompts analyzed in this study are provided below to give a general overview of the score information for the comparison groups. Table 1 presents overall means and standard deviations of the raw essay and ELA scores for the Indo-European and Non-Indo-European language groups. Standardized mean differences (i.e., Cohen's *d*) between the two groups are also provided.

As shown in Table 1, the ELA and observed essay scores were higher for the Indo-European language group than for the non-Indo-European language group. The standardized mean difference in the MELAB essay score, 0.49, would be viewed as a "small" effect size

(Cohen, 1988). The standardized mean difference in the ELA score, 0.24, may also be viewed as a "small" effect size.

Table 1. Means, Standard Deviations, and Standardized Group Mean Differences (Cohen's *d*) for Essay and ELA Scores for Indo-European (IE) and Non-Indo-European (NIE) Language Groups

Variable/Language group	Sample size	Mean	SD	*d*
MELAB essay score				
IE group	575	77.39	6.82	0.49
NIE groups	1694	74.40	5.87	
English language ability				
IE group	575	0.18	1.03	0.24
NIE groups	1694	-0.06	0.97	

Table 2 presents overall means and standard deviations of the observed essay and ELA scores for the male and female examinees. Standardized mean differences (i.e., Cohen's *d*) between the two groups are also provided.

As shown in Table 2 below, the observed essay scores were higher for the female group than for the male group. The standardized mean difference in the MELAB essay score, 0.12, would be viewed as a "small" effect size according to Cohen's (1988) standard. The standardized mean difference in the ELA score was extremely small, as indicated by the almost identical group means.

Table 2. Means, Standard Deviations, and Standardized Group Mean Differences (Cohen's *d*) for Essay and ELA Scores for Male and Female Examinees

Variable/Gender	Sample size	Mean	SD	*d*
MELAB essay score				
Male	686	74.65	6.46	0.12
Female	1583	75.39	6.16	
English language ability				
Male	686	0.00	1.91	0.01
Female	1583	0.00	1.75	

T-Tests

As a preliminary analysis, independent samples t-tests were conducted to compare the means of the comparison groups. The results are shown in Tables 3 and 4 for the different language and gender groups, respectively.

Table 3. Means, Standard deviations, and Independent Samples T-Test Results by Language Group

	IE group		NIE group			
Prompt	Mean	SD	Mean	SD	*t*	*d*
18	77.58 (n=57)	6.24	74.31 (n=109)	5.61	3.42*	0.55
25	78.15 (n=67)	5.87	76.04 (n=166)	5.50	2.59*	0.37
46	77.44 (n=81)	5.44	75.49 (n=284)	5.00	3.03*	0.38
48	77.37 (n=67)	5.97	74.99 (n=154)	5.72	2.80*	0.40
49	79.88 (n=78)	7.85	75.05 (n=236)	5.82	5.80*	0.75
51	76.77 (n=31)	7.93	74.30 (n=117)	5.98	1.90	0.38
52	75.93 (n=42)	6.77	73.44 (n=152)	6.13	2.27*	0.39
54	74.62 (n=47)	8.08	73.40 (n=172)	6.90	1.03	0.17
55	76.42 (n=31)	6.66	73.32 (n=130)	5.23	2.81*	0.55
56	77.19 (n=74)	7.11	72.49 (n=174)	6.14	5.25*	0.72

*$p < 0.05$, two-tailed.

Table 4. Means, Standard deviations, and Independent Samples T-Test Results by Gender

	Male group		Female group			
Prompt	Mean	SD	Mean	SD	*t*	*d*
18	75.52 (n=44)	6.45	75.40 (n=122)	5.89	0.11	0.02
25	76.41 (n=72)	5.36	76.75 (n=161)	5.83	-0.41	0.06
46	74.69 (n=87)	5.47	76.31 (n=278)	5.01	-2.58*	0.31
48	75.75 (n=81)	5.66	75.69 (n=140)	6.03	0.07	0.01
49	74.91 (n=112)	7.20	76.99 (n=202)	6.31	-2.66*	0.31
51	75.87 (n=39)	6.48	74.44 (n=109)	6.48	1.18	0.22
52	73.70 (n=47)	7.70	74.07 (n=147)	5.86	-0.34	0.06
54	75.51 (n=41)	7.65	73.93 (n=178)	7.04	-1.14	0.19
55	74.34 (n=93)	5.87	73.32 (n=68)	5.31	1.13	0.18
56	72.13 (n=70)	6.63	74.58 (n=178)	6.73	-2.59*	0.36

*$p < 0.05$, two-tailed.

As shown in Table 3 above, for the two different language groups, significant differences in the means were found in eight out of the ten prompts analyzed. In terms of the standardized group mean score difference index (i.e., Cohen's *d*), one, five, and four prompts had effect sizes considered small, medium, and large, respectively (Cohen, 1988). For the gender comparison groups, as presented in Table 4, significant differences in the means were found in three out of the ten prompts analyzed. Six prompts (18, 25, 48, 52, 54, and 55) had small effects, and the four remaining prompts (46, 49, 51, and 56) had medium effects. It should be noted that simple differences in mean scores on an item across different examinee subgroups is not evidence of bias or unfairness. In some cases, examinees from two different groups may actually differ in the ability of interest, and differences in item performance are to be expected. These results are referred to as *item impact* (Clauser & Mazor, 1998). In fact, the real fairness issue should be the extent to which DIF is present in any of the MELAB writing prompts. The results obtained from the logistic regression DIF method are provided below.

Logistic Regression DIF

As described in the method section, a three-step modeling process based on ordinal logistic regression (Zumbo, 1999) was used as the main method of analysis. For example, the language group DIF results for prompt 25 can be summarized as follows:

Step 1. Model with ELA only: $\chi^2(1) = 112.96$, R-squared = 0.1102
Step 2. Uniform DIF: $\chi^2(1) = 113.97$, R-squared = 0.1112
Step 3. Uniform and Nonuniform DIF: $\chi^2(1) = 114.52$, R-squared = 0.1118

Examining the difference between steps 1 and 3 above for the DIF test, the resulting statistics are $\chi^2(2) = 1.57$, p = 0.4568, and R-squared = 0.0016. Given a nonsignificant p-value of 0.4568 (p > 0.01), this prompt is not demonstrating DIF across different language groups. Besides, the R^2 effect size measure ($R^2\Delta$), which provides information about the "practical" significance of DIF (Zumbo, 1999), is negligible even by Jodoin and Gierl's (2001) conservative standard ($R^2\Delta < 0.035$).

The results of the logistic regression DIF analysis for the language and gender comparison groups are summarized in Tables 5 and 6, respectively. As shown in Tables 5 and 6 below, although a few prompts analyzed had statistically significant group differences ($p < 0.01$), the R^2 effect sizes (i.e., $R^2\Delta$ in Tables 5 and 6) were far too small for any prompt to be classified as having a group effect either across different language groups or across gender.

Table 5. Model Comparisons for the 10 MELAB Prompts across Different Language Groups

Prompt	Model	-2 Log likelihood	R^2	Model comparison	Difference			
					χ^2	df	*p*	$R^2\Delta$
18	1	83.61	0.1125	1 vs. 3	8.38	2	0.0151	0.0113
	2	91.98	0.1238	1 vs. 2	8.37	1	0.0038*	0.0113
	3	91.99	0.1238	2 vs. 3	0.01	1	0.9161	0.0000
25	1	112.96	0.1102	1 vs. 3	1.57	2	0.4568	0.0016
	2	113.97	0.1112	1 vs. 2	1.01	1	0.3151	0.0001
	3	114.52	0.1118	2 vs. 3	0.56	1	0.4551	0.0006
46	1	150.93	0.0987	1 vs. 3	4.98	2	0.0831	0.0032
	2	153.86	0.1006	1 vs. 2	2.92	1	0.0872	0.0019
	3	155.91	0.1019	2 vs. 3	2.05	1	0.1521	0.0013
48	1	105.79	0.1069	1 vs. 3	3.38	2	0.1842	0.0034
	2	109.03	0.1102	1 vs. 2	3.24	1	0.0720	0.0033
	3	109.17	0.1103	2 vs. 3	0.15	1	0.7015	0.0001
49	1	174.48	0.1185	1 vs. 3	14.95	2	0.0006*	0.0101
	2	181.74	0.1234	1 vs. 2	7.28	1	0.0070*	0.0049
	3	189.43	0.1286	2 vs. 3	7.68	1	0.0056*	0.0052
51	1	76.86	0.1148	1 vs. 3	12.11	2	0.0023*	0.0181
	2	87.75	0.1311	1 vs. 2	10.89	1	0.0010*	0.0163
	3	88.97	0.1329	2 vs. 3	1.21	1	0.2707	0.0018
52	1	95.69	0.1085	1 vs. 3	10.06	2	0.0065*	0.0114
	2	105.34	0.1194	1 vs. 2	9.64	1	0.0019*	0.0109

Prompt	Model	-2 Log likelihood	R^2	Model comparison	Difference			
					χ^2	df	p	$R^2\Delta$
	3	105.75	0.1199	2 vs. 3	0.41	1	0.5209	0.0005
54	1	104.62	0.1024	1 vs. 3	5.11	2	0.0778	0.0049
	2	109.13	0.1068	1 vs. 2	4.51	1	0.0337	0.0044
	3	109.73	0.1073	2 vs. 3	0.60	1	0.4388	0.0005
55	1	75.29	0.1070	1 vs. 3	6.33	2	0.0422	0.0009
	2	80.64	0.1146	1 vs. 2	5.35	1	0.0207	0.0076
	3	81.62	0.1160	2 vs. 3	0.98	1	0.3212	0.0014
56	1	156.95	0.1336	1 vs. 3	16.08	2	0.0003*	0.0137
	2	168.22	0.1432	1 vs. 2	11.27	1	0.0008*	0.0096
	3	173.03	0.1473	2 vs. 3	4.81	1	0.0283	0.0041

*$p < 0.01$, Model 1 predictor: ELA, Model 2 predictors: ELA, Language group, Model 3 predictors: ELA, Language group, ELA*Language group.

Note.
Model comparison 1 vs. 3: Uniform + Nonuniform DIF
Model comparison 1 vs. 2: Uniform DIF only
Model comparison 2 vs. 3: Nonuniform DIF only

Effect size measure (Jodoin & Gierl, 2001)

$R^2\Delta < 0.035$: "negligible" effect
$0.035 < R^2\Delta < 0.070$: "moderate" effect
$R^2\Delta > 0.070$: "large" effect

Table 6. Model Comparisons for the 10 MELAB Prompts across Gender

Prompt	Model	-2 Log likelihood	R^2	Model comparison	Difference			
					χ^2	df	p	$R^2\Delta$
18	1	83.61	0.1125	1 vs. 3	0.45	2	0.7968	0.0006
	2	84.06	0.1131	1 vs. 2	0.45	1	0.5004	0.0006
	3	84.06	0.1131	2 vs. 3	0.00	1	0.9868	0.0000
25	1	112.96	0.1102	1 vs. 3	2.36	2	0.3075	0.0023
	2	115.31	0.1125	1 vs. 2	2.35	1	0.1251	0.0023
	3	115.31	0.1125	2 vs. 3	0.01	1	0.9341	0.0000
46	1	150.93	0.0987	1 vs. 3	14.15	2	0.0008*	0.0092
	2	163.62	0.1070	1 vs. 2	12.69	1	0.0004*	0.0083
	3	165.08	0.1079	2 vs. 3	1.46	1	0.2270	0.0009
48	1	105.79	0.1069	1 vs. 3	1.59	2	0.4505	0.0016
	2	106.76	0.1079	1 vs. 2	0.97	1	0.3247	0.0010
	3	107.39	0.1085	2 vs. 3	0.63	1	0.4292	0.0006
49	1	174.48	0.1185	1 vs. 3	11.65	2	0.0030*	0.0079
	2	181.07	0.1229	1 vs. 2	6.59	1	0.0102	0.0044
	3	186.12	0.1264	2 vs. 3	5.06	1	0.0245	0.0035

Prompt	Model	-2 Log likelihood	R^2	Model comparison	Difference			
					χ^2	df	p	$R^2\Delta$
51	1	76.86	0.1148	1 vs. 3	2.91	2	0.2335	0.0044
	2	79.47	0.1187	1 vs. 2	2.61	1	0.1064	0.0039
	3	79.77	0.1192	2 vs. 3	0.30	1	0.5823	0.0005
52	1	95.69	0.1085	1 vs. 3	4.62	2	0.0994	0.0052
	2	96.54	0.1095	1 vs. 2	0.85	1	0.3564	0.0010
	3	100.31	0.1137	2 vs. 3	3.77	1	0.0523	0.0042
54	1	104.62	0.1024	1 vs. 3	10.64	2	0.0049*	0.0104
	2	109.65	0.1073	1 vs. 2	5.02	1	0.0250	0.0049
	3	115.26	0.1128	2 vs. 3	5.61	1	0.0178	0.0055
55	1	75.29	0.1070	1 vs. 3	1.91	2	0.3840	0.0028
	2	75.97	0.1080	1 vs. 2	0.68	1	0.4108	0.0010
	3	77.20	0.1098	2 vs. 3	1.24	1	0.2659	0.0018
56	1	156.95	0.1336	1 vs. 3	20.68	2	0.0000*	0.0177
	2	176.31	0.1501	1 vs. 2	19.36	1	0.0000*	0.0165
	3	177.63	0.1513	2 vs. 3	1.32	1	0.2507	0.0012

*$p < 0.01$, Model 1 predictor: ELA, Model 2 predictors: ELA, Language group, Model 3 predictors: ELA, Language group, ELA*Language group.

Note.
Model comparison 1 vs. 3: Uniform + Nonuniform DIF
Model comparison 1 vs. 2: Uniform DIF only
Model comparison 2 vs. 3: Nonuniform DIF only

Effect size measure (Jodoin & Gierl, 2001)

$R^2\Delta < 0.035$: "negligible" effect
$0.035 < R^2\Delta < 0.070$: "moderate" effect
$R^2\Delta > 0.070$: "large" effect

Discussion and Conclusion

The number of tasks that can be feasibly administered in direct assessments of writing, such as in essay tests, is usually small because such formats of assessment require extended responses and are time consuming to administer and score (Powers & Fowles, 1999). Often only one writing prompt is administered to each examinee, as in the MELAB composition part. Under such circumstances, it is very important to ensure that each prompt is as fair as possible to examinee subgroups.

The primary purpose of this study was to investigate DIF across different language and gender groups on the MELAB essay test, using the logistic regression method. A three-step modeling process based on ordinal logistic regression (Zumbo, 1999) employed in this study seemed to be efficient in investigating simultaneously both uniform and nonuniform group effects related to native languages and gender. In the tradition of logistic regression DIF tests, in this study, the term DIF is synonymous with the simultaneous test of uniform and

nonuniform DIF with a two-degree-of-freedom Chi-squared test. The statistical significant tests were supplemented with a measure of the corresponding effect size via R-squared.

The results of this study showed that although a few prompts analyzed were initially flagged through the three-step logistic regression method due to the significant uniform and/or nonuniform group effects, their effect sizes were far too small for any of them to be classified as DIF essay items. That is, the essay score differences between the different groups compared in this study seem to be due to "item impact" rather than "group difference" attributable to a construct irrelevant factor inherent in writing prompts. A clear distinction is usually made between "item impact" and "DIF" in the item bias literature (Clause & Mazor, 1998; Penfield & Lam, 2000; Zumbo, 1999). "Item impact" may be present when examinees from different groups have different probabilities of success on an item, because examinees from these groups do actually differ in the ability of interest. In such circumstances, group differences in examinee performance on the item are to be expected because of "true" differences between the groups in the underlying ability being measured by the item.

In general, it was believed that the logistic regression procedures worked well in this study. However, some limitations of the current study should be noted. One limitation was the ELA variable used in this study may not be an ideal matching variable. A better matching variable would have been a measure similar to the free-response writing prompts being studied. Since each examinee responds to a single writing prompt in the MELAB composition part, there was no similar matching variable available. The use of a multiple-choice measure such as ELA as a matching variable assumes that if examinees have high ELA measured by the two parts of the test as a whole, they should perform well overall on the essays and vice versa. An important question is whether similar effect sizes might have been obtained if a more direct measure of writing had been used as a matching variable. It may be possible to conduct research that would answer this question in the future.

A second limitation is related to the sample size used in the present study. In the ordinal logistic regression procedure, accurate estimation of parameters depends on a healthy sample size in each score category. Considering the sample size used, therefore, the results of this study should be interpreted with caution. An important question is whether similar results could have been obtained if larger sample sizes per group (e.g., $n > 200$) had been used in the logistic regression procedure.

In conclusion, even though "judgmental" methods that rely on expert judges' opinions can be used to select potentially biased prompts, it is a relatively subjective procedure. Thus, it is also recommended that "statistical" techniques (i.e., DIF procedures) be routinely implemented to identify differentially functioning items for various comparison groups. Prompt developers can benefit from routinely identifying prompts through statistical quality control and then reviewing those that are identified as extreme.

Acknowledgement

I would like to extend my sincere gratitude to the English Language Institute of the University of Michigan for funding this research project.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chang, H., Mazzeo, J., & Roussos, L. A. (1995). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure (ETS Research Report No. 95–5). Princeton, NJ: Educational Testing Service.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31–44.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Schmitt, A. J. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.

English Language Institute, University of Michigan. (2003). *MELAB technical manual*. Ann Arbor, MI: English Language Institute, University of Michigan.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315–332.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

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

Lee, Y., Breland, H., & Muraki, E. (April, 2002). Comparability of TOEFL CBT essay prompts for different native language groups. Paper presented at the annual conference of National Council on Measurement in Education, New Orleans, LA.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*, 107–122.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple group partial credit model. *Journal of Educational Measurement*, *36*, 217–232.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5–15.

Powers, D. E., & Fowles, M. E. (1999). Test-takers' judgments of essay prompts: Perceptions and performance. *Educational Assessment, 6*, 3–22.

Sheppard, L. A. (1982). Definition of bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore, MA: Johns Hopkins University.

Spray, J., & Miller, T. (1994). *Identifying non-uniform DIF in polytomously scored test items.* (Research Report No. 93–1). Iowa City, IA: American College Testing Program.
Weigle, S. C. (2000). Test review: The Michigan English language assessment battery (MELAB). *Language Testing, 17*, 449–455.
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, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.
Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance task. *Journal of Educational Measurement, 30*, 233–251.

Appendix A

MELAB Composition Rating Scale

97 Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence level) structures, accurate morphological (word forms) control. Organization is appropriate and effective, and there is excellent control of connection. There is a wide range of appropriately used vocabulary. Spelling and punctuation appear error free.

93 Topic is fully and complexly developed. Flexible use of a wide range of syntactic structures. Morphological control is nearly always accurate. Organization is well controlled and appropriate to the material, and the writing is well connected. Vocabulary is broad and appropriately used. Spelling and punctuation errors are not distracting.

87 Topic is well developed, with acknowledgement of its complexity. Varied syntactic structures are used with some flexibility, and there is good morphological control. Organization is controlled and generally appropriate to the material, and there are few problems with connection. Vocabulary is broad and usually used appropriately. Spelling and punctuation errors are not distracting.

83 Topic is generally clearly and completely developed, with at least some acknowledgement of its complexity. Both simple and complex syntactic structures are generally adequately used; there is adequate morphological control. Organization is controlled and shows some appropriacy to the material, and connection is usually adequate. Vocabulary use shows some flexibility, and is usually appropriate. Spelling and punctuation errors are sometimes distracting.

77 Topic is developed clearly but not completely and without acknowledging its complexity. Both simple and complex syntactic structures are present; in some "77" essays these are cautiously and accurately used while in others there is more fluency and less accuracy. Morphological control is inconsistent. Organization is generally controlled, while connection is sometimes absent or unsuccessful. Vocabulary is adequate, but may sometimes be inappropriately used. Spelling and punctuation errors are sometimes distracting.

73 Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some "73" essays both simple and complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Organization is partially controlled, while connection is often absent or unsuccessful. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Spelling and punctuation errors are sometimes distracting.

67 Topic development is present but restricted, and often incomplete or unclear. Simple syntactic structures dominate, with many errors; complex syntactic structures, if present, are not controlled. Lacks morphological control. Organization, when apparent, is poorly

controlled, and little or no connection is apparent. Narrow and simple vocabulary usually approximates meaning but is often inappropriately used. Spelling and punctuation errors are often distracting.

63 Contains little sign of topic development. Simple syntactic structures are present, but with many errors; lacks morphological control. There is little or no organization, and no connection apparent. Narrow and simple vocabulary inhibits communication, and spelling and punctuation errors often cause serious interference.

57 Often extremely short; contains only fragmentary communication about the topic. There is little syntactic or morphological control, and no organization or connection are apparent. Vocabulary is highly restricted and inaccurately used. Spelling is often indecipherable and punctuation is missing or appears random.

53 Extremely short, usually about 40 words or less; communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control, and no apparent organization or connection. Vocabulary is extremely restricted and repetitively used. Spelling is often indecipherable and punctuation is missing or appears random.

N.O.T. (Not On Topic) indicates a composition **written on a topic completely different from any of those assigned**; it does not indicate that a writer has merely digressed from or misinterpreted a topic. N.O.T. compositions often appear prepared and memorized. They are not assigned scores or codes.