Issue in Language Teaching (ILT), Vol. 8, No. 1, 33-61, June 2019

Investigating Gender and Major DIF in the Iranian National University Entrance Exam Using Multiple-Indicators Multiple-Causes Structural Equation Modelling

Hamdollah Ravand*
Associate Professor, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

Tahereh Firoozi
Ph.D. Candidate, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

Gholamreza Rohani
Assistant Professor, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

Abstract
The generalizability aspect of construct validity, as proposed by Messick (1989), requires that a test measure the same trait across different samples from the same population. Differential Item Functioning (DIF) analysis is a key component in the fairness evaluation of educational tests. The university entrance exam for Master of English programs (hereinafter referred to as MEUEE) at Iranian state universities is a very high-stakes test whose fairness is a promising line of research. The current study explored gender and major DIF in the general English (GE) section of the MEUEE using multiple-indicators multiple-causes (MIMIC) structural equation modeling. The data of all the test takers (n = 21,642) who took the GE section of the MEUEE in 2012 were analyzed with Mplus. To determine whether an item is flagged for DIF, both practical and statistical significance were considered. The results indicated that 12 items were flagged for DIF in terms of statistical significance. However, only 5 items showed practical significance. The items flagged for DIF alert test developers and users to potential sources of construct-irrelevant variance in the test scores that may call into question comparison of the test-takers' performances, especially when the tests are used for selection purposes.

Keywords: Differential item functioning, multiple-indicators multiple-causes structural equation modelling, university entrance examination

***Corresponding author's email:** ravand@vru.ac.ir

INTRODUCTION

Fairness in the accountability era has been regarded as the essence of any valid measurement instrument. When a measure is used to make decisions that have serious consequences for stakeholders, its fairness needs due attention. To decide whether the interpretation and use of a test are equally fair to different subgroups, technical issues of measurement, such as differential item functioning (DIF), come into play (Newton & Shaw, 2014).

DIF occurs when group membership (e.g., male vs. female) of the test takers, rather than their knowledge of the construct being measured, decides how they perform on any item of a given test. The presence of DIF can adversely impact both selection and classification inferences. Hence, various studies have investigated DIF in high-stakes tests such as TOEFL (e.g., Bailey, 1999; Stricker & Rock, 2008; Wall & Horák, 2008), and SAT (e.g., Bridgeman & Wendler, 1991; Curley & Schmitt, 1993; Kanarek, 1988). However, DIF studies on high-stakes tests in general and university entrance examinations (UEEs) in particular are in short supply in the Iranian context. Therefore, to justify the validity of tests as life-changing as the UEEs, more validation studies in general and DIF studies, in particular, are needed.

Bringing more evidence to the fairness of UEEs by considering the generalizability and consequential aspects of Messick's (1995) validity framework, the current study attempted to explore gender and major DIF in the items of the general English (GE) section of the MEUEE. To this end, we have applied the multiple-indicators multiple-causes (MIMIC) structural equation modeling (SEM) approach to DIF suggested by Woods (2009).

The research questions guiding the current study are:

1. How do the items of the GE section of the MEUEE function in terms of the candidates' gender?
2. How do the items of the GE section of the MEUEE function in terms of the test-takers' Bachelor's field of study?

LITERATURE REVIEW

DIF studies can be located within the *generalizability* and *consequential* aspects of the validity framework suggested by Messick (1996). The generalizability aspect concerns the principle of invariance which is assumed to be the essence of validity argument in educational assessment. According to the invariance principle, items of a test are supposed to be a sample of all possible items which could be included in the test and persons who take a given test are a sample of the population of all possible test takers. The item and person invariance need due attention when comparing test results across different groups of items and persons. According to Messick (1995, p. 746) "the consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short-and long-term". The unintended outcomes may be positive, like improving educational systems, or negative, as the source of item bias or DIF. That is why researchers have tried to address the validity of some high-stakes tests such as Test of English as a Foreign Language (TOFEL), Scholastic Aptitude Test (SAT), Graduate Record Examination (GRE), and International English Language Testing System (IELTS) through DIF Analysis.

Various studies have investigated DIF in high-stakes tests such as TOEFL (e.g., Bailey, 1999; Stricker & Rock, 2008; Wall & Horák, 2008), and SAT (e.g., Bridgeman & Wendler, 1991; Curley & Schmitt, 1993; Kanarek, 1988;). For example, Chen and Henning (1985) conducted one of the earliest DIF investigations on language proficiency tests. In their study, they examined DIF on the English as a Second Language Placement Examination (ESLPE) for examinees with different language backgrounds. In another study, Ryan and Bachman (1992) used the Mantel–Haenszel procedure for the detection of items that functioned differentially across Indo-European and Non-Indo-European language groups on the First Certificate of English (FCE) and TOEFL. In the Iranian context, Birjandi and Amini (2007) investigated the DIF items and the possible causes of DIF in the IELTS' listening and reading sections. In their study, DIF items were

identified using item response theory (IRT), likelihood ratio approach (LR) and Mantel-Haenszel statistical procedures. Gender DIF has also been explored in a variety of high-stakes tests using diverse DIF detection methods (Carlton & Harris, 1992; Curley & Schmitt, 1993; Gafni, 1991; Lawrence, Curley & McHale, 1988; Lawrence & Curley, 1989; O'Neill & McPeek, 1993; O'Neill, McPeek & Wild, 1993; Scheuneman & Gerritz, 1990).

DIF studies on high-stakes tests in general and university entrance examinations (UEEs) in particular are in short supply in the Iranian context. High-stakes tests used for selection purposes in Iran are developed and administered by the Measurement Organization (MO). Among these tests are UEEs that screen test-takers into Bachelor, Master, and Ph.D. levels of different majors at both state and non-state universities (formerly just at state universities). Validity studies on the UEEs are far and few between. There have been some sporadic DIF studies on the UEE at Bachelor's level (e.g., Barati, Ketabi & Ahmadi, 2006).

As a part of a larger validation study using the Rasch model, Ravand and Firoozi (2016) investigated DIF in the university entrance exam for the candidates who seek admission into master's English programs at the Iranian state universities (MEUEE). In another recent study, Ahmadi and Darabi (2016) investigated gender DIF in the Ph.D. entrance examination for the applicants into English programs using item response theory (IRT) and logistic regression approaches. However, to justify the result of tests as life-changing as the UEEs, more validation studies in general and DIF studies, in particular, are needed.

Considering the generalizability and consequential aspects of Messick's (1995) validity framework, the current study attempted to explore gender and major DIF in the items of the general English (GE) section of the MEUEE. The significance of the present study is threefold: (1) Practically, the study is a quest to bring evidence for the validity of the MEUEE which would help test constructors to take measures to avoid item bias in future versions of the test, (2) Methodologically, the present study uses multiple

indicators multiple causes (MIMIC) structural equation model (SEM) to detect DIF, which enjoys the robustness and flexibilities of the SEM. MIMIC models were popularized as a DIF detection method by Muthen (1989). Woods, Oltmanns and Turkheimer (2009) enumerated the following merits for the MIMIC as a DIF detection method: It uses latent ability estimates as a matching variable which is more accurate than observed summed scores since their estimation involves taking measurement error into account. Multidimensional items (i.e., items which measure multiple factors) are easily modeled. They can be implemented with general-purpose SEM software such as Mplus (Muthen & Muthen, 2007) which can handle categorical item response data. With MIMIC models DIF can be readily tested for more than two groups simultaneously and use more than one covariate as a controlling variable. Last but not the least, it has been shown (Woods, 2009) that the MIMIC model yields more accurate DIF estimates when the size of the focal group is very small (e.g., 25-100). (3) Pedagogically, demonstrating the step by step application of the MIMIC model in DIF testing makes the study an educational source. To take advantage of the robustness and flexibilities of the rather recent MIMIC approach to DIF detection, studies are required to walk the readers through the procedures required to implement the approach.

METHOD

Data Collection

The data of the present study came from the responses of all the applicants (n = 21642) to the general English (GE) section of the MEUEE in 2012. The MEUEE is developed and administered by the Iranian Measurement Organization to screen applicants into English Teaching, English Literature, and Translation Studies programs at the Master of Arts (MA) level in Iran. The MEUEE measured both GE and content knowledge. The present study used the data of the GE section which measured English proficiency and was composed of 60 multiple-choice (MC) items: Structure (10 items),

vocabulary (20 items), cloze test (10 items), and reading comprehension (20 items). The structure and vocabulary sections measured the structural and lexical knowledge in standard written English. The cloze test consisted of one gapped passage of 180 words of which 10 had been removed selectively from the text. The last section of the GE section included three passages of nontechnical reading materials, two of which were followed by 7 questions and the other passage by 6 questions. The reading comprehension items measured test takers' ability to understand the gist, main idea, and logical argument and recognize details, writers' opinion, attitude, and purpose. Table 1 present the distribution of participants' gender and field of study.

Table 1: Gender and major distribution of the subjects

Gender	%	Number	Major	%	Number
Female	73%	15832	English	88%	19037
Male	27%	5810	Non-English	12%	2605
Total		21,642	Total		21,642

As Table 1 shows, about two-thirds of the subjects were females and their fields of study at Bachelor's level were mainly the English language.

Data Analysis

The data were analyzed using Mplus (version 6.11; Muthen & Muthen 2007). Since the data were categorical, models were fitted to the data using the robust maximum likelihood (MLR) estimator.

The data analysis procedure of the study was conducted in the following steps:

a) Exploratory factor analysis (EFA)
b) Confirmatory Factor Analysis (CFA)
c) Testing the items for DIF through the MIMIC model

To test for DIF, the following steps were taken for each of the three factors (i.e., grammar, vocabulary, and reading): First, the DIF-free items were identified, then all the other items, technically referred to as *studied*

items, were tested for DIF one at a time. Lastly, a final model was specified wherein all the items flagged for DIF in the previous step were regressed on the grouping variables (here gender and major). To decide whether an item should be flagged for DIF both chi-square difference test and the alternative fit indices (AFI) suggested by Cheung and Rensvold (2002) were used. According to the AFI, comparative fit index (CFI) values in the nested models are compared with the CFI value of a baseline model. According to Cheung and Rensvold, $\Delta CFI \geq 0.01$ indicates model non-invariance and flags the respective item for DIF. Conventionally, items flagged for DIF according to the chi-square difference test are said to have statistical significance, and those flagged according to AFI are said to have practical significance.

Results

To see whether the items of the test would cluster under the different sections as they appear on the test, EFA was run. In order to decide on how many factors to retain, Cattell's (1966) scree plot and simple structure methods were used. In the scree plot method the number of factors to retain is the number of nodes above the point of inflection in the plot and according to the simple structure method, factors with high loadings (≥ 0.5) from at least three to five items and cross-loading < 0.32 on the other factors (Costello & Osborne, 2005; Tabachnick & Fidell, 2001) are retained. According to the eigenvalues or the Kaiser Criterion, there nine factors that explained the clustering of the data. However, since the extraction of this many factors was not compatible with the current understanding of language proficiency, Scree Plots of the eigenvalues were consulted. Both the scree plot (Figure 1) and simple structure methods suggested three factors accounted for the variance in the item-level data.

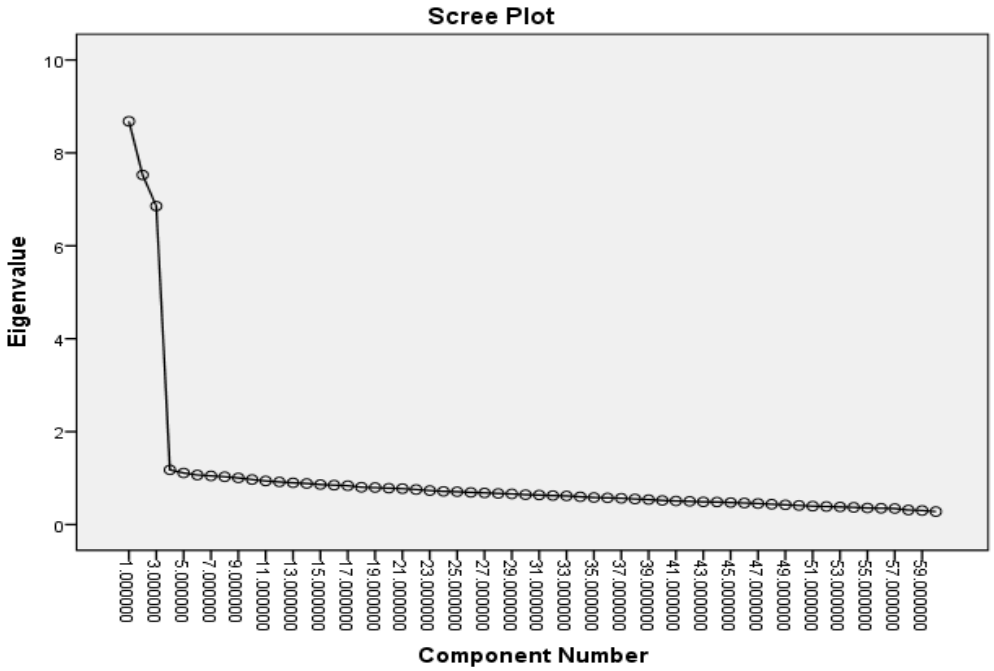

Figure 1: Scree plot

In the next step, the factor structure obtained through EFA was tested using the CFA model as shown in Figure 2.

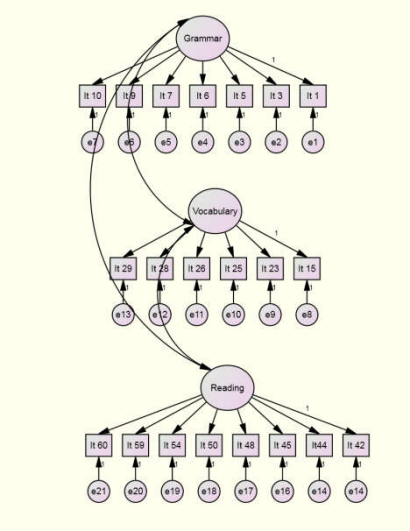

Figure 2: The underlying construct of the GE section of the MEUEE

A CFA model can be evaluated at two levels: (1) At a macro level, global fit indices such as chi-square test of model fit (x^2), Comparative Fit Index (CFI; Bentler, 1990), Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), and Root Mean Square Error of Approximation (RMSEA) are checked. For the chi-square, a non-significant value suggests a good fit. However, since x^2 is sensitive to sample size, the other global fit indices are preferred. CFI and TLI values ≥ 0.95 indicate a well-fitting model (Hu & Bentler, 1999) and an RMSEA value ≤ 0.05 indicates a good fit. As Table 2 shows, except for a significant chi-square value, all the indices show a good fit of the model to the data.

Table 2: Model fit of CFA model

Model Fit	
Chi-square	*1826.334
CFI	0.950
TLI	0.954
RMSEA	0.02

(2) At a micro level, the significance and plausibility of the individual parameters in the model are checked. In assessing individual parameter estimates there are two aspects of concern (Byrne, 2012): (a) The appropriateness of the estimates. In a measurement model, viable estimated values should exhibit the correct sign and fall within the admissible range. Parameter estimates taken from covariance or correlation matrices that are not positive definite, as well as estimates exhibiting out-of-range values such as correlations > 1.00 and negative variances, exemplify unacceptable estimated values; (b) statistical significance of the estimates. As Table 3 shows, all the factor loadings are positive, within range, and statistically significant, indicating a good fit of the model to the data.

Table 3: Parameter fit of CFA model

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
F1 BY				
It 1	1.000	0.000	999.000	999.000
It 3	0.704	0.023	31.297	0.000
It 5	0.868	0.023	38.442	0.000
It 6	0.946	0.022	42.896	0.000
It 7	0.825	0.021	38.575	0.000
It 9	0.910	0.021	43.093	0.000
It 10	0.684	0.024	28.521	0.000
F2 BY				
It 15	1.000	0.000	999.000	999.000
It 23	2.428	0.189	12.845	0.000
It 25	1.855	0.146	12.725	0.000
It 26	0.404	0.094	4.289	0.000
It 28	2.253	0.174	12.915	0.000
It 29	0.565	0.077	7.338	0.000
F3 BY				
It 42	1.000	0.000	999.000	999.000
It 44	2.000	0.151	13.202	0.000
It 45	2.571	0.187	13.735	0.000
It 46	1.203	0.108	11.169	0.000
It 48	2.692	0.200	13.487	0.000
It 50	2.566	0.191	13.403	0.000
It 54	2.262	0.170	13.314	0.000
It 59	1.836	0.142	12.883	0.000
It 60	2.533	0.188	13.478	0.000

Note. F1: Grammar; F2: Vocabulary; F3: Reading
It: Item; SE: standard error

The second column in Table 3 lists unstandardized factor loadings. For identification purposes, the loading of one of the items on each factor has been fixed to 1. All the remaining free estimated parameters show reasonable positive values as well as statistical significance. Both the overall model fit indices and the parameters of the model suggest that a three-factor model as specified in Figure 2 shows a good fit.

Specifying the Hypothesized MIMIC Model

To address the first and second research questions, three hypothesized MIMIC models for different dimensions (factors) of the MEUEE were tested in three steps: identifying DIF free items, testing each item for DIF, and specifying the final model. To make comparisons, first, a baseline model for each of the factors is specified. Figure 3 is the hypothesized baseline model for grammar. The same model was specified for reading and vocabulary, which are not presented here in the interest of space. In these baseline models, the hypothesis is that the MEUEE is measurement invariant (DIF-free) with no direct paths from the grouping variable to the items.

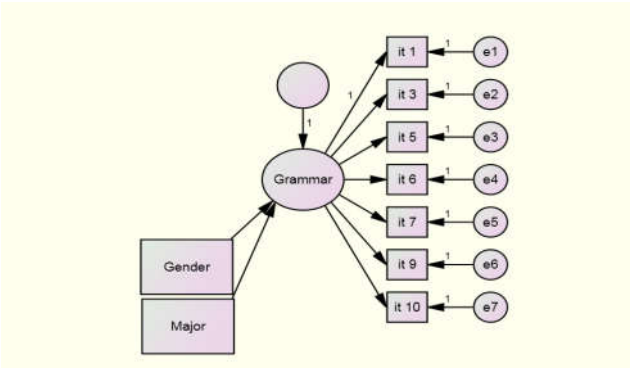

Figure 3: The hypothesized MIMIC model for grammar

Identifying DIF Free Items: In this step, each item is tested for DIF while other items are presumed DIF-free. The aim is approached by regressing one item at a time on all the grouping variables (Figure 4).

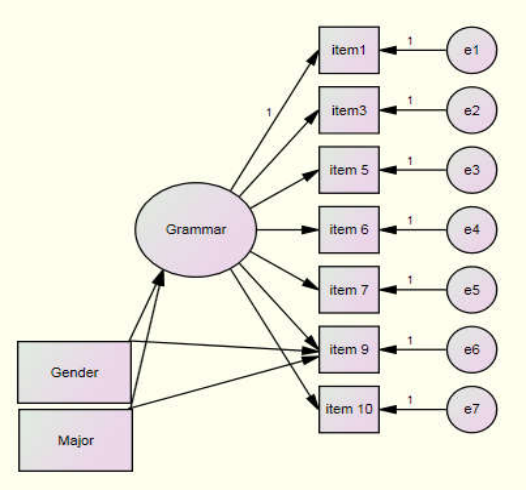

Figure 4: Identifying DIF free items in the grammar section

Non-significant estimates of the regression paths and the factor loadings ≥ 0.5 are indicators of *anchored* or DIF-free items (Woods, 2009). According to the estimated values represented in Table 4, Items 1, 3, 6, 7, 23, 44, and 59 are major DIF-free and Items 5, 6, 7, 10, 25, 44, 45, 48, 54, 59, and 60 are gender DIF-free. The regression path estimates on these items are non-significant (P> 0.05) and the factor loadings are ≥ 0.5. Therefore, these items were assigned as anchored items and the rest of the items were assigned as studied items that needed to be tested individually for DIF in the next step.

Table 4: Stage One Item Parameters

Items	F BY	ON	Estimate	S.E	Est./S.E.	Two-Tailed P-Value
IT 1	1.000	GENDER	-0.037	0.027	-1.347	0.178
		MAJOR	0.072	0.019	3.807	0.000
IT 3	0.751	GENDER	0.016	0.027	0.584	0.559
		MAJOR	0.051	0.020	2.466	0.014

IT 5	0.923	GENDER	0.089	0.027	3.306	0.001
		MAJOR	-0.033	0.020	-1.658	0.097
IT 6	1.007	GENDER	-0.025	0.026	-0.937	0.349
		MAJOR	0.004	0.019	0.208	0.836
IT 7	0.849	GENDER	-0.007	0.027	-0.267	0.790
		MAJOR	0.001	0.019	0.049	0.961
IT 9	0.943	GENDER	-0.099	0.027	-3.703	0.000
		MAJOR	-0.062	0.019	-3.265	0.001
IT 10	0.692	GENDER	0.093	0.028	3.317	0.001
		MAJOR	-0.038	0.021	-1.841	0.066
IT 15	1.000	GENDER	0.211	0.032	6.654	0.000
		MAJOR	0.146	0.022	6.765	0.000
IT 23	0.645	GENDER	-0.056	0.035	-1.604	0.109
		MAJOR	-0.058	0.023	-2.579	0.010
IT 25	0.868	GENDER	-0.081	0.032	-2.564	0.010
		MAJOR	-0.021	0.021	-0.999	0.318
IT 26	0.385	GENDER	0.094	0.039	2.382	0.017
		MAJOR	-0.059	0.025	-2.347	0.019
IT 28	1.110	GENDER	-0.194	0.037	-5.177	0.000
		MAJOR	-0.052	0.022	-2.354	0.019
IT 29	0.439	GENDER	0.094	0.032	2.962	0.003
		MAJOR	-0.050	0.021	-2.362	0.018
IT 42	1.000	GENDER	0.054	0.027	2.008	0.045
		MAJOR	0.071	0.019	3.668	0.000
IT 44	1.777	GENDER	-0.009	0.027	-0.339	0.735
		MAJOR	0.015	0.019	0.794	0.427
IT 45	2.136	GENDER	0.087	0.028	3.119	0.002
		MAJOR	-0.011	0.020	-0.547	0.584
IT 48	1.746	GENDER	0.069	0.027	2.572	0.010
		MAJOR	-0.017	0.020	-0.866	0.387
IT 50	1.959	GENDER	0.142	0.027	5.263	0.000
		MAJOR	-0.072	0.020	-3.673	0.000
IT 54	2.243	GENDER	-0.220	0.028	-7.758	0.000
		MAJOR	0.002	0.020	0.091	0.928
IT 59	1.634	GENDER	0.042	0.027	1.527	0.127
		MAJOR	0.018	0.019	0.950	0.342
IT 60	2.613	GENDER	-0.153	0.033	-4.628	0.000
		MAJOR	0.026	0.022	1.146	0.252

Note. **IT: Item**

Testing Each Item for DIF: All the studied items from the previous stage of the analysis were tested individually for DIF using both chi-square-based likelihood ratio test (LRT) and alternative fit indices (AFIs). To test each item for DIF, first, a full model wherein all the studied items are supposed

to have DIF (all the items are regressed on the grouping variable) is specified. Then the studied items are removed one at a time. When the first studied item is removed, the fit of the model is compared against that of the full model however fit of each succeeding model is compared against that of the model immediately preceding it. In the current study, since just one regression path is removed at a time in the nested models (there is just one degree of freedom difference between each model and its preceding model), a chi-square difference, denoted as $\Delta\chi^2 > 3.841$ shows the fit of the model with the parameter removed is worsened compared to its preceding model wherein the given parameter was included. Therefore, the removed item is flagged for DIF.

Table 5: Testing each item for DIF

	$\Delta\chi^2$	
Items	**ON GENDER**	**ON MAJOR**
IT 1		10.521 *
IT 3		6.791 *
IT 5	11.260 *	
IT 9	7.492 *	4.650 *
IT 10	12.010 *	1.937
IT 15	13.034 *	19.801 *
IT 23		3.628
IT 25	1.009	
IT 26	8.100 *	4.808 *
IT 28	1.427	0.194
IT 29	11.567 *	4.191 *
IT 42	2.418	12.669 *
IT 45	3.735	
IT 48	3.278	
IT 50	10.381 *	12.732 *
IT 54	31.598 *	
IT 60	13.715 *	

As Table 5 shows, the chi-square difference test flagged for gender DIF the following items: 5, 9, 10, 15, 26, 29, 50, 54, and 60 and major DIF

items: 1, 3, 9, 15, 26, 29, 42, and 50. As it is clear from Table 6, in testing items 10, 23, 28 for major DIF, and items 25, 28, 42, 45, 48 for gender DIF individually, the invariance of nested and full models indicated non-zero but insignificant DIF for these items; however, they were flagged for DIF in the previous stage of the analysis. Hence, adhering to the law of parsimony, these non-significant items were removed in the next step.

Since chi-square and as a result, chi-square difference tests, are sensitive to sample size, in large sample sizes the power to detect even trivial differences leads to a lack of invariance in two models. Therefore, to identify practical DIF, the AFIs suggested by Cheung and Rensvold (2002) were also consulted. To approach this end, we compared the CFI values in the nested models with the CFI value of the baseline model (Table 6). According to Cheung and Rensvold, $\Delta CFI \geq 0.01$ indicates model non-invariance and flags the given item for DIF. Table 7 summarizes the CFI indices of the studied items.

Table 6: Checking each item for DIF (Alternative fit indices)

		CFI	
Baseline Grammar		**0.990**	
	On G&M	**On G**	**On M**
Item 1 removed			0.989
Item 3 removed			0.990
Item 5 removed		0.989	
Item 9 removed	0.989	0.990	0.990
Item 10 removed	0.989	0.989	0.990
Baseline vocabulary		**0.976**	
Item 15 removed	0.964	0.971	0.969
Item 23 removed			0.976
Item 25 removed		0.977	
Item 26 removed	0.973	0.973	0.975
Item 28 removed	0.977	0.977	0.977
Item 29 removed	0.972	0.972	0.976
Baseline Reading		**0.974**	
Item 42 removed	0.971	0.974	0.972
Item 45 removed		0.973	
Item 48 removed		0.974	
Item 50 removed	0.970	0.972	0.972
Item 54 removed		0.968	
Item 60 removed		0.971	

Note. G: Gender; M: Major

According to Table 6, Items 1 and 15 show major DIF and Items 5, 10, and 54 show gender DIF.

Specifying the Final MIMIC Models: In this stage, the fit of the final models which were constructed according to the second step was tested. In each model, only items that showed significant DIF were regressed on the grouping variables. The factors were also regressed on the grouping variables. The final models provide estimates of factor loadings, group mean differences on the factor and DIF effects.

Table 7: The general fit of the final MIMIC model to the data

Model Fit			
	Grammar	**Vocabulary**	**Reading**
Chi-square	116.897*	70.170*	166.345*
CFI	0.990	0.977	0.974
TLI	0.983	0.952	0.959
RMSEA	0.021	0.020	0.021

Table 7 shows the general fit of the models. The chi-square values are significant but the alternative fit indices are within the acceptable range: CFI and TLI ≥ 0.095, and RMSEA < 0.05.

Table 8: Parameters of the final MIMIC model to the data

Grammar Model						
Items	**F1**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 1	1.000*			0.073*	Gender	0.216*
It 3	0.757*			0.056*	Major	0.130*
It 5	0.922*		0.088*			
It 6	1.015*					
It 7	0.856*					
It 9	0.962*		-0.069*	-0.038		
It 10	0.689*		0.095*			
Vocabulary Model						
Items	**F2**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 15	1.000*		0.244*	0.137*	Gender	0.259*

It 23	0.647*				Major	0.159*
It 25	0.879*					
It 26	0.346*		0.124*	-0.047		
It 28	1.135*					
It 29	0.401*		0.127*	-0.034		
			Reading Model			
Items	**F2**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 42	1.000*		0.029	0.067*	Gender	0.029*
It 44	1.771*				Major	0.058*
It 45	0.879*					
It 48	0.346*					
It 50	1.135*		0.076*	-0.066*		
It 54	0.401*		-0.234*			
It 59	1.632*					
It 60	2.613*		-0.188*			

As Table 8 shows all the factor loadings (Column 2) and group mean differences in the factor (Column 6) are significant ($P < 0.05$). Columns 3and 4 show the estimate of the items' regression paths on the grouping variables or DIF effect. Inspection of the final model results shows that the regression coefficients for paths from major to Items 26, and 29 and gender to Item 42, which were flagged for DIF in the previous stage, are not statistically significant in the final models. Therefore, the models were modified by omitting the paths from the respective grouping variables to Items 26, 29, and 42. Furthermore, the negative signs of the regression paths from gender to Items 9, 54, and 60 and from major to Items 26 and 29 indicate that the level of latent variable required for the focal groups (males and non-English majors) to correctly answer these items was lower than that needed for the corresponding reference groups (females and English major).

Re-specification of the final models involves the omission of three regression paths: Item 42 on gender and Items 9, 26, and 29 on major. As Table 8 shows, the DIF effects of these items were non-significant, so the paths were candidates for deletion. The result of the SEM analysis of modified models is summarized in Tables 9 and 10.

Table 10: The parameter fit of the modified final MIMIC model

Grammar Model						
Items	**F1**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 1	1.000*		0.082*	Gender	0.216*	
It 3	0.758*		0.063*	Major	0.121*	
It 5	0.924*	0.088*				
It 6	1.016*					
It 7	0.857*					
It 9	0.959*	-0.068*				
It 10	0.690*	0.096*				
Vocabulary Model						
Items	**F2**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 15	1.000*	0.243*	0.145*	Gender	0.259*	
It 23	0.648*			Major	0.158*	
It 25	0.878*					
It 26	0.348*	0.124*				
It 28	1.134*					
It 29	0.402*	0.127*				
Reading Model						
Items	**F2**	**BY**	**ON Gender**	**On Major**	**F**	**ON**
It 42	1.000*		0.067 *	Gender	0.032*	
It 44	1.765*			Major	0.058*	
It 45	2.127*					
It 48	1.740*					
It 50	1.979*	0.072*	-0.066*			
It 54	2.237*	-0.239*				
It 59	1.627*					
It 60	2.613*	-0.193*				

Table 9: General fit of the modified final MIMIC models to the data

Model Fit			
	Grammar	**Vocabulary**	**Reading**
Chi-square	116.897*	74.917**	166.621*
CFI	0.990	0.976	0.974
TLI	0.983	0.957	0.960
RMSEA	0.021	0.019	0.021

Reviewing goodness-of-fit statistics related to the models, one can see that little improvement in both overall fit and parameter estimates is observed. The three respecified models are preferred over the previous ones

on the following grounds: (1) The respecified models are more parsimonious, and (2) all the indices are significant.

In summary, according to the chi-square difference test, 12 items in the GE section of the MEUEE showed statistically significant DIF across gender and major. From among these items, 7 showed gender DIF, 3 major DIF, and 2 showed both gender and major DIF. Due to the large sample size of the study, besides the results obtained based on the chi-square difference test, the results of AFIs were also consulted. According to the results of AFI, Items 5, 9, 10, 15, and 54 showed practical DIF in terms of participants' gender and Items 1, 9, 10, and 15 in terms of major. According to Table 11, about 50% of grammar items, 15% of vocabulary items, and 20% of reading items showed statistically significant gender and major DIF.

Table 11: Statistically significant DIF items

	Grammar	**Vocabulary**	**Reading**
Female	4	3	1
Male	1	0	2
English Major	2	1	1
Non-English Major	0	0	1
Total	5	3	4
Percent	50%	15%	20%

However, as Table 12 shows only 15% of the grammar items, 5% of the vocabulary items and, 5% of the reading items showed practical DIF. Comparison of the results in Tables 12 and 13 shows that the numbers of practical DIF items are much less than statistically significant DIF items.

Table 12: Practically significant DIF Items

	Grammar	**Vocabulary**	**Reading**
Gender	3	0	1
Major	3	1	0
Total[a]	3	1	1
Percent	15%	5%	5%

Note[a]. Items with both gender and major DIF were considered as one item.

DISCUSSION

The present study investigated gender and major DIF among the items of the GE section of the MEUEE using the MIMIC model. Besides providing evidence as to the performance of the respective items, the current study presented an illustration of the MIMIC approach to DIF detection. The results showed that about 41% of all DIF items were in the grammar section. One reason why the grammar section showed by far the highest number of DIF items may be the fact that the test was a speed test, and the grammar section appears the first followed by the vocabulary and reading comprehension sections on the GE section of the MEUEE. Thus, if the first come first served to say is at work on educational tests, most of the test takers must have tried the grammar items and more variability where observed across genders and majors on these items.

The grammar part was mostly favored by females. This finding is supported by other studies in the literature (e.g., Barati & Ahmadi, 2010; Karami, 2011). According to Karami (2011), 19 out of 100 items on the University of Tehran English Proficiency Test (UTEPT) showed statistically significant DIF but only three items had practical significance and all were grammar items. Furthermore, a study conducted by Barati and Ahmadi (2010) revealed that about 59. % of the grammar items in the UEE for the applicants into Bachelor's English programs showed gender DIF in favor of females. The result is in line with Carlton and Harris (1992) and O'Neill and McPeek (1993) who concluded that females would outperform males on abstract concepts. The grammar section of the MEUEE features decontextualized items that measure structural knowledge such as word order in standard written English.

In the current study, the majority of statistically significant DIF items in the reading section were in favor of males. The content of the first two passages was more scientific than the third one. The passage related to Item 50 which showed DIF in favor of females was about "the harmful effect of waste" while, the passage related to Items 54 and 60 which showed DIF in

favor of males was social—"the changes in family structure". As to topic familiarity, it is expected that females perform better on social topics than a matched group of males and overall, the expectation is that the number of gender DIF items in reading section be higher than those of the other sections (Curely & Schmitt, 1993; Lawrence et al. 1988; Maller, 2001; Scheuneman & Gerritz 1990; Wild & McPeek 1986). These two expectations were not met in the current study, so the result cannot be explained by the topic familiarity issue. However, one reason for the above contradictions may be that MEUEE is regarded as a speed test. Since the reading section is at the end of the test, the participants missed most of the items in this part due to lack of time. Hence, the participants' response function varies less in this section than the grammar part as the first part of the test.

As to the items flagged for major DIF, Item 50 was in favor of non-English major participants. To focus the source or "why question" of DIF, the content of the item was inspected qualitatively. The finding may be justified in two ways: (1) topic familiarity, and (2) de-contextualization. This reading comprehension item tests scanning strategy as one of the reading skills. The item is referenced to a science passage about "the harmful effect of waste". As the majority of non-English major participants studied technical major as Bachelor's level, they were more familiar with science-related topics. Furthermore, no contextual clues were needed to correctly answer this item. Put it another way, this item was de-contextualized in a way that by having a little background or topical knowledge about "termination of wastes", the participants could easily answer the item correctly without referring to the passage.

Items 1, 3, 15 and 42 favored English majors. Although Item 1 tested the word order as the grammatical knowledge, the familiarity of participants with the prefabricated pattern, "in its widest sense", was essential to give the correct answer to the item. Item 15, in the vocabulary section, tapped into the use of collocation "elapsed time". English major participants had more target language use situations to get familiar with the use besides the usage

of English than non-English majors. In other words, English majors had more years of schooling in English fields than non-English majors to gain knowledge of language functions, prefabricated patterns, collocations which are beyond decontextualized words and sentences. Because the factors causing DIF in these items are not irrelevant to the construct underlying the test, they may be considered as impact rather than bias.

Items 3 and 42 tested grammatical word order in written English and inferential knowledge as a reading strategy, respectively. Analyzing the content qualitatively, I could find no clues of bias in these items. However, one reason for this finding may be that English major participants are more risk-takers to make good guesses than non-English major counterparts. According to Oxford and Nyikos's (1989) the year and the field of schooling influence the strategy preferences and orientations of mental behavior. Oxford and Nyikos's (1989) claimed that the students who studied a language for at least four years are better risk-takers than students with fewer years of study. Furthermore, they argued that humanities or social science majors used independent strategies more than technical majors did. According to this finding, English major Participants performed better than non-English major counterparts on these two items because they used more resourceful and independent strategies, such as elaborating sentences or listing related words.

CONCLUSION AND IMPLICATIONS

The results of this study provide evidence to the test constructors on how the items of MEUEE function in terms of participants' gender and field of study. The highlighted DIF items make the test developers aware of the existence of DIF that may differentially affect the performance of individuals with the same ability level. Due to the significant amount of resources allotted to MEUEE development and the seriousness of the consequences of the test interpretation and use, more care on the part of test constructors is required to be more accountable to the stakeholders.

Although there has been a lot of progress in devising DIF detection methods, there has been little progress in understanding the cause of DIF. Understanding why DIF occurs requires follow-up substantive analysis. Most DIF studies have focused have done away with the substantive analysis. Even some of these studies (e.g., Cole 1981; Linn 1986; Plake 1980; Tittle, 1982) have argued that attempts to find reasons for DIF items do not usually come up with results that can be relied on. However, there have been some attempts to frameworks for DIF analysis. Roussos and Stout (1996) proposed a two-step approach that was intended to bridge the gap between the statistical and substantive DIF analysis. In the first stage through qualitative analysis sources of difficulty of test items are identified. The sources of difficulty could be some strategies, attributes, etc. which may be construct-relevant or construct-irrelevant. In the second stage, DIF items are detected through statistical analysis. Combining information obtained from the two stages can help find DIF items and the cause of DIF in these items. When an item is flagged for DIF in the statistical analysis stage, without information from the qualitative analysis stage, one might jump to the conclusion the respective item is biased. As it was mentioned before, for DIF to occur two conditions should be met: (1) There are source of difficulty involved in answering any given item other than the primary attribute the item is intended to measure, (2) Performance of the focal and reference groups with similar levels of the primary attribute differ significantly on at least one of the other attributes. Even when the above two conditions are met, one cannot claim DIF unless the other attributes are construct-irrelevant. Thus, the results obtained in the present study and other similar DIF studies should not be interpreted before amassing sufficient information from the qualitative analysis of the test items as to what knowledge sources or attributes are involved in getting each item right.

Mere statistical analysis for DIF detection is not enough. Once the DIF items are identified, investigating the source of DIF by combining qualitative and quantitative information is crucial. To identify knowledge sources required to get the items on the GE section of the UEUEE expert

judgment, think-aloud protocol analysis of the test takers and empirical methods such as diagnostic classification models can be used. Using both statistical DIF analysis and substantive DIF analysis, a distinction should be made between items with the adverse DIF (i.e., where a construct irrelevant dimension is the source of differential performance) and those with the benign DIF (i.e., where a construct-relevant dimension is the source of differential performance). Benign DIF contributes to construct validity whereas adverse DIF is a threat to construct validity.

In the current study, a version of the MIMIC model which is capable of detecting uniform DIF was employed. Future studies can use a different version of the MIMIC model (Woods, 2011) to study both uniform and non-uniform DIF in the MEUEE items. With a focus on methodological aspects, future studies can apply different DIF detection methods including multiple groups to study DIF in the MEUEE and compare the results.

REFERENCES

Ahmadi, A., & Darabi, A. (2016). Gender differential item functioning on a national field-specific test: The case of Ph.D. entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research, 4*(1), 63-82.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bailey, K. (1999). *Washback in language testing (TOEFL Monograph Series 15)*. Princeton, NJ: Educational Testing Service.

Barati, H., & Ahmadi, A., R. (2010). Gender-based DIF across the subject area: A study of the Iranian national university entrance exam. *Journal of Teaching Language Skills, 2*(3), 1-26.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high stakes tests: The effect of field of study. *IJAL*, *19*(2), 27-42.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 237-246.

Birjandi, P., & Amini, M. (2007). Differential item functioning (test bias) analysis paradigm across manifest and latent examinee groups (on the construct validity of IELTS). *Journal of Human Sciences, 8*(2), 1-20.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics classes. *Journal of Educational Psychology*, *83*(2), 275-284.

Byrne, B. M. (2012). *Multivariate applications series. Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge/Taylor & Francis Group.

Carlton, S. T., & Harris, A. M. (1992). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*(10), 1067-1077.

Costello, A. B., & Osborne, J. W. (2005). Exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(7), 1-9.

Curley, W. E., & Schmitt, A. P. (1993). Revising SAT-verbal items to eliminate differential item functioning. *ETS Research Report Series*, *1993*(2), i–18.

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach.* New York, NY: Routledge.

Gafni, N. (1991). *Differential item functioning: performance by sex on reading comprehension tests. ERIC document no. ED 331844.* Rockville, MD: Educational Resources Information Center.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Kanarek, E. A. (1988). *Gender differences in freshman performance and their relationship to use of the SAT in admissions.* Paper read at the annual meeting of the Regional Association for Institutional Research. October, at Providence, RI

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 27-38.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492.

Lawrence, I. M., & Curley, W. E. (1989). *Differential item function for males and females on SAT-Verbal Reading subscore items: Follow-up study. Educational Testing Service Research Report*. Princeton, NJ: Educational Testing Service.

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT verbal reading subscore items. Report No. 88-4*. New York, NY: College Entrance Examination Board.

Linn, M. C., De Benedictis, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "i don't know" really mean? *Journal of Research in Science Teaching, 24*(3), 267-278.

Linn, M. C., & Petersen, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. In J. Hyde & M. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 67-101), Baltimore, MD: Johns Hopkins University.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292-326.

Maller, S. J. (2001). Differential item functioning in the Wisc-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 5D, 741-749.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational and Behavioral Statistics, 10*(2), 121-132.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage Publication.

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

O'Neill, K. A., McPeek, W. M., & Wild, C. L. (1993). *Differential item functioning on the graduate management admission test. ETS Research Report 93–35*. Princeton, NJ: Educational Testing Service.

Osborne, J. W. (2013). Data Cleaning Basics: Best practices in dealing with extreme scores. *Newborn and Infant Nursing Reviews, 10*(1), 37-43.

Oxford, R., & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *The Modern Language Journal, 73*(3), 291-300.

Pae, T. I., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475-496.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40*(2), 397-404.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing, 6*(1), 1-18.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355-371

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12-29 .

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131.

Stricker, L. J.*, &* Rock, D. A. (2008). *Factor structure of the TOEFL Internet-Based Test across subgroups (TOEFL iBT Research Report 07)*. Princeton, NJ: Educational Testing Service.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*, Needham Heights, MA: Allyn & Bacon.

Tittle, C. K. (1982). Use of judgemental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 31-63). Baltimore, MD: The Johns Hopkins University Press.

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. doi: 10.1007/bf02291170

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe—Phase 2, coping with change.* Princeton, NJ: Educational Testing Service.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non normal variables: Problems and remedies. In R. H. Hoyle (Ed.) *Structural equation modelling: Concepts, issues, and applications* (pp. 56-75), Thousand Oaks, CA: Sage.

Wild, C., & McPeek, W. (1986). *Performance of the Mantel-Haenszel statistic in identifying differentially functioning items.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Willingham, W. W., Cole, N. S., Lewis, C., & Leung, S. W. (1997). Test performance. In W. W. W. N. S. Cole (Ed.), *Gender and fair assessment* (pp. 55-126). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Willse, J. T., Goodman, J. T., Allen, N., & Klaric, J. (2008). Using structural equation modeling to examine group differences in assessment booklet designs with sparse data. *Applied Measurement in Education*, 21(3), 253-272.

Woods, C., Oltmanns, T., & Turkheimer, E. (2009). Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*(4), 320-330.

Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behavioral Research, 44*(1), 1-27. doi: 10.1080/00273170802620121

Woods, C. M. (2011). DIF testing for ordinal items with poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement*, 35, 145-164.