An Assessment of Performance Differentials by Gender and Educational Level in ELI Placement Test

SLS 490, Fall 2005

Thamana Lekprichakul

An assessment of performance differentials by gender and educational level in ELI placement test

Introduction

Each year, the University of Hawaii at Manoa admits approximately 1,000 immigrant and international students of which 20%[1] are required to take the English Language Institute's (ELI) placement test. The purposes of the ELI placement test are two folds: (a) to determine the proficiency levels of the students' academic English and (b) to identify areas in which additional ELI classes may be needed to facilitate their studies at the University of Hawaii. The ELI placement tests consist of five parts, i.e., essay[2], dictation, academic listening, cloze, and reading comprehension tests. The essay is intended to test academic writing ability, dictation and listening tests are to test academic listening comprehension skills; and the reading comprehension which includes vocabulary as sub-tests and cloze tests are designed to assess academic reading ability. Note, however, that students are not tested for their spoken language proficiency.

The primary purpose of this study is to examine to see whether the ELI placement tests are equally fair to students of every possible sub-group. Here, test fairness is narrowly defined as being absent from testing bias. This study is conducted in response to the ELI department's policy in ensuring fair and valid assessment of students' academic English. The report proceeds in the following order. It starts off by defining testing bias and scope of study. The next section describes methods of bias analysis. Data descriptions and results are to follow. The report will touch upon reliability and validity issues before summing up the paper and discussing limitation

[1] The estimate is provided by Kenton Harsch, the current Assistant Director of the ELI, during a private meeting.
[2] ELI essay test is for graduate and special unclassified exchange students only. Undergraduates take a different writing test called Manoā writing placement exam to place students in different writing classes. For greater details about the ELI placement test, please visit ELI's website: http://www.hawaii.edu/eli/.

of the study and issues for further study. Potential differential items functioning (DIF) are reported in the appendix.

Definition and Scope of Study

What is testing bias? Test bias is said to exist when the following key conditions are met. Firstly, there must be performance differentials between the focal group and the norm group either at the test item or score level. Secondly, the performance difference is attributable to feature of the test that is "***not*** relevant to what is being measured"[3] (emphasis by the author). The former is a necessary condition and the latter a sufficient condition. Identifying performance differentials between groups as bias requires that both conditions be met. The construct-irrelevant source of variation that systematically helps or hurts the scores of a group of examinees over the other is some time referred to as measurement bias.

Scope of study

Determining performance differentials between groups is a straight forward statistical matter. It is, however, difficult to determine whether such performance differences of an item in question can be traced to factors that are construct-relevant. Often time, experts in language, in general, or language testing, in particular, are needed to conduct a thorough sensitivity review[4] of differentially functioning test items to make such determination. Due to time constraint, this study limits its scope to analyze performance differentials of designated group of interest and identify items that are potentially biased. Test review is not the subject of this investigation.

Source of bias

Test bias can originate from many sources. Gender, geography, native language, nationality, educational background, and social classes are just a few examples. In this paper, only gender and educational enrollment level (undergraduates vs. graduates) will be examined. Other variables of potential interest such as language background, nationality, students' major or planned major will not be considered in this study because data is not yet available for bias analysis. It is worth noting that bias analysis and differential item functioning (DIF) analysis are used interchangeably in this paper.

---

[3] ALTE (1998) p. 136 cited in Brown (2005) p. 246.
[4] Source: Unknown author, http://siop.org/_Principles/pages31to34.pdf. (12/5/2005). SIOP stands for the Society for Industrial and Organizational Psychology.

Method of Bias Analysis

How testing bias is analyzed is tied closely to how it is defined. Brown (2005) listed two approaches to empirically define test bias: i.e., the legal and statistical definitions. The legal definition is based on item difficulty (ID) or item facility differential (IF)[5]. In 1984, the court ruling in the Golden Rule Insurance, Co vs. Mathias case defined a bias item as any item with IF differential of 0.15 or higher. An alternative legal definition of test bias is given by the Equal Employment Opportunity Commission (EEOC) which imposes 80% rule for the selection rate of the protected minority versus that of the majority group. This EEOC's 80% rule implies that test item is considered biased if the IF differential is greater than 0.20. On the other hand, the statistical definition does not rely on value judgments which use a pre-set threshold IF differential value to define biased item. Rather, the statistical definition is based on and varies with data. Any statistically significant difference in group mean indicates potential bias. There are two statistical approaches to bias analysis, i.e., the item-response-theory (IRT) based and non-item-response-theory based statistical analysis. Each has its own strength and weakness. For this exploratory purpose, the simple non-IRT method suffices.

Mean comparisons play key roles in detecting test bias. One way analysis of variance (ANOVA) and *t*-test will be extensively used to test whether any significant difference between group means at total score, sub-total score and item levels are present. Legal definitions will be used to identify DIF items. To cross examine the DIF results, Kunnan's (1990) method of outlier detection will be used to identify DIF items. The central idea of Kunnan's method is to fit a regression line on a scatter plot of the IF between groups coupled with 95% confidence interval. Any observation outside the 95% confidence band is considered a DIF item.

[5] Here, item difficulty and item facility differential are used interchangeably.

Data

The data for this bias analysis are from the ELI's placement test in Fall 2004. There are 202 examinees that complete all four tests: academic listening, dictation, cloze and reading tests. Each sub-test has 50 items (k) except the listening test that has only 49 items. The scores are non-weighted with possible maximum raw scores of 199. Table 1 shows summary statistics of the scores.

Table 1

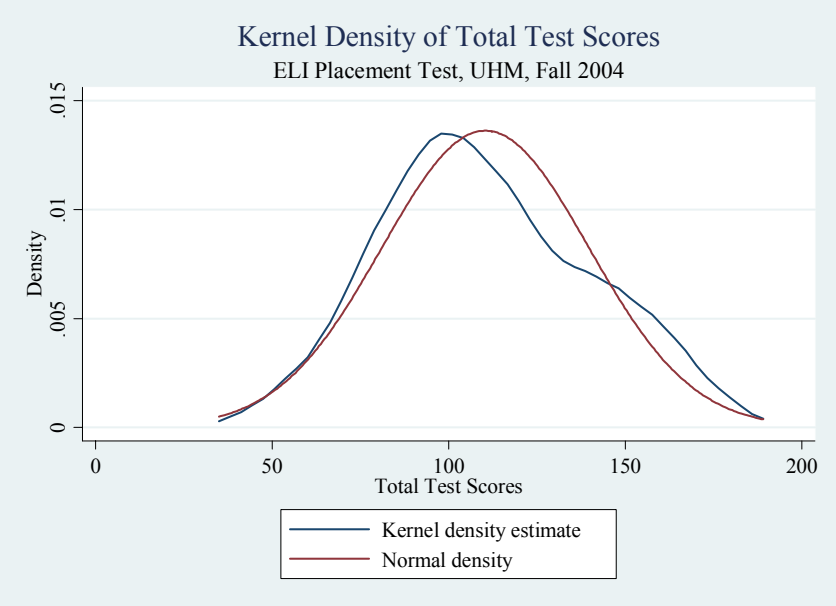
Descriptive statistics of test scores, ELI placement test, UHM Fall 2004

Statistics	Total	Listening	Cloze	Dictation	Reading
Mean	110.42	28.44	24.59	28.41	29.19
Median	107	28	25	27	29
Midpoint	112.00	29.00	24.00	28.50	29.00
Max	180	47	42	50	49
Min	44	11	6	7	9
N	202	207	208	211	206
k	199	49	50	50	50
Std. Dev.	29.29	7.23	8.09	10.71	8.63
Skewness	0.25	0.12	-0.02	0.25	-0.03
Kurtosis	2.44	2.42	2.40	2.14	2.27

Note that the skewness statistics are approximately zero indicating that the distributions of scores are fairly symmetrical. However, the positive and sizable kurtosis statistics indicates that the score distributions are of the "leptokurtic" type with relatively large tails. Since the placement is a norm reference test (NRT), it is important to determine whether test scores are normally distributed. Figure 1 and 2 shows observed distributions of scores in relation to the normal distribution curves. Test statistics using χ^2 and Shapiro-Wilk's w statistics indicate that the distribution of total score deviates from normal and the cause of this deviation is the dictation score. Note, however, that the χ^2 and w statistics disagree on whether the distribution of reading score is normal. While the χ^2 statistics rejects the null hypothesis that reading score is normally

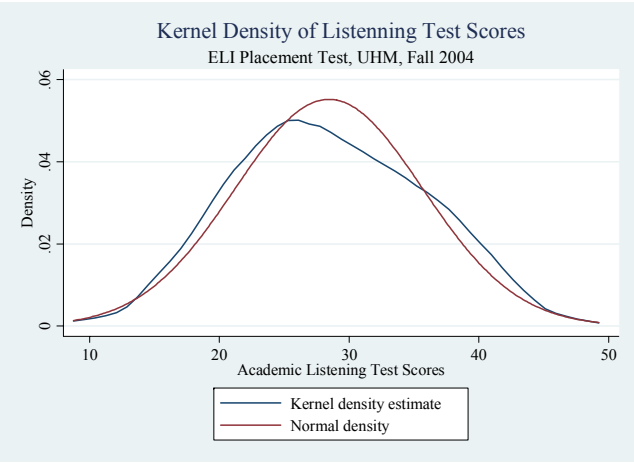
distributed with $p < 0.01$, the *w* statistics indicates that there is no evidence of deviation from normal distribution ($p > 0.05$). Upon visual inspection of the distribution, I agree with *w* statistics that the distribution of reading score is possibly normal.

At issue are what effect and its effect size these non-normal distributions have in distributing students to different level of proficiency level including effects on correlation coefficients and reliability estimates. The non-normal distribution of total score has no effect on distribution of students because the ELI primarily focuses *not* on the total score but rather on the scores of sub-test to identify areas where students need help. On the other hand, a possible inadvertent impact of the non-normal distribution of the dictation score may be that disproportionately high percentage of students, in relation to what would have been otherwise if listening placement were based on listening test instead, may have been exempted or placed in high level of ELI's listening related classes.

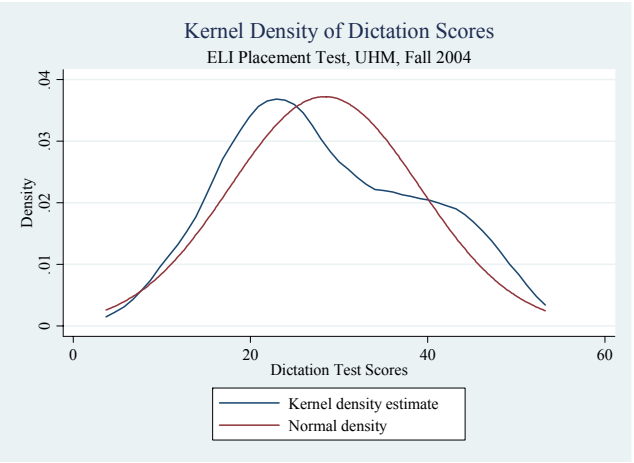


Total Score: Non-Normal Distribution ($p<0.05$)

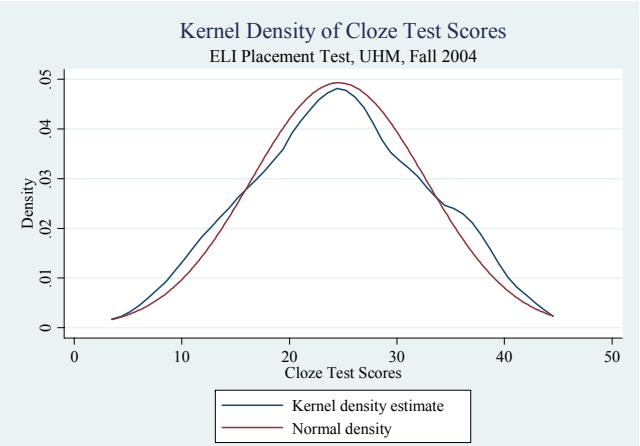
Figure 1 The observed probability density distribution (PDF) of the total score plotted against the normal distribution curves using kernel smoothing technique. Normality test using χ^2 statistics and Shapiro-Wilk's *w* statistics were used to test the null hypotheses of normal distributions.



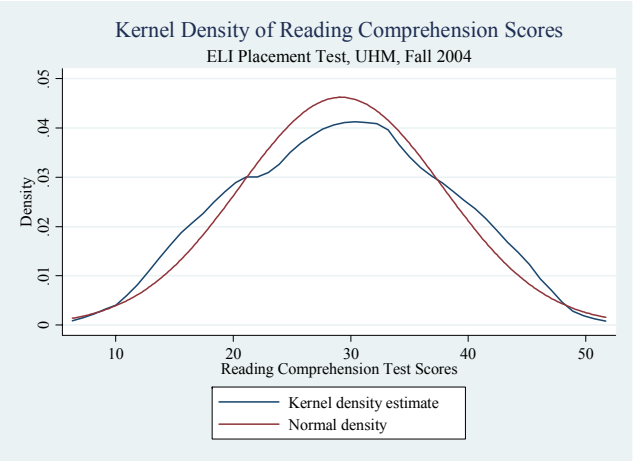
Listening: Normal Distribution



Dictation: Non-Normal Distribution (p<0.01)



Cloze: Normal Distribution



Reading: Possibly Normal Distribution

Figure 2 The observed probability density distributions (PDF) of each score are plotted against the normal distribution curves using kernel smoothing technique. Normality test using χ^2 statistics and Shapiro-Wilk's w statistics were used to test the null hypotheses of normal distributions.

Table 2

Number and percentage distribution of examinees by gender and education level

Designated group of interest	*N*	Percent
	Gender	
Female	130	61.0
Male	83	39.0
Total	213	100.0
	Education level	
Undergraduates	177	79.7
Graduates	45	20.3
Total	222	100.0

Table 2 shows the number and distribution of examinees by gender and educational enrollment level. Three out of every five test takers are female, a sex ratio of 1.5 to 1. The majority of examinees are enrolling at the undergraduate level and one out of every five students taking tests are enrolling at the graduate level.

Result

Bias Analysis of Total Score

Figure 3 shows a bar graph average total score by gender and education level. The average scores of female and male examinees are almost identical, i.e., 110.4 and 110.5 respectively. For education variable, the average score of the undergraduate students is slightly higher than that of the graduate counterpart, i.e., 111.1 vs. 107.9. ANOVA indicates that there are no statistically significant differences between the group means by either gender or education level. From the total score perspective, there is no evidence to suggest neither that the ELI placement test unfairly advantages or disadvantages examinees of either gender nor that it systematically helps or hurts undergraduate or graduate students.

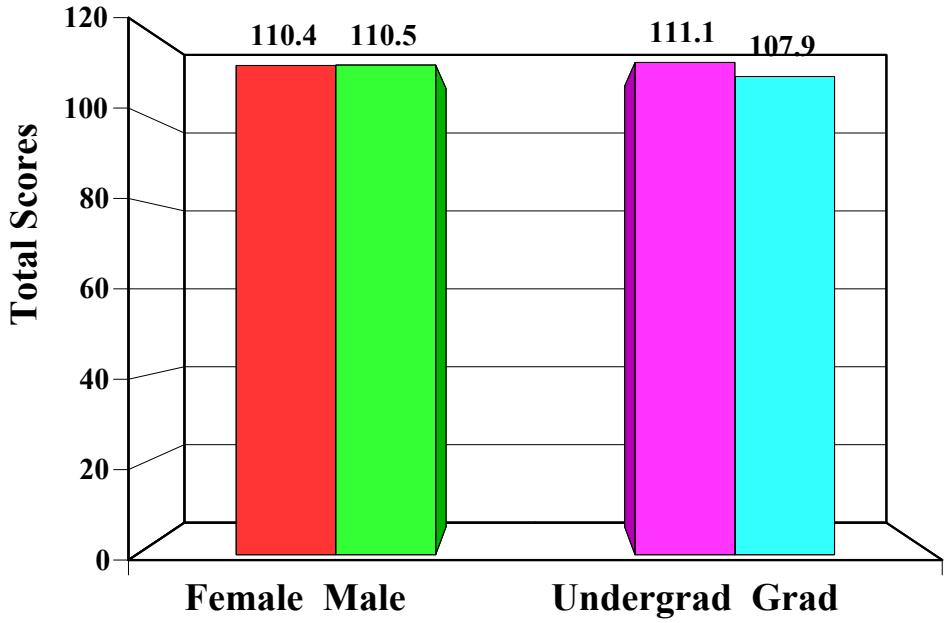


Figure 3 Average total scores by gender and education level

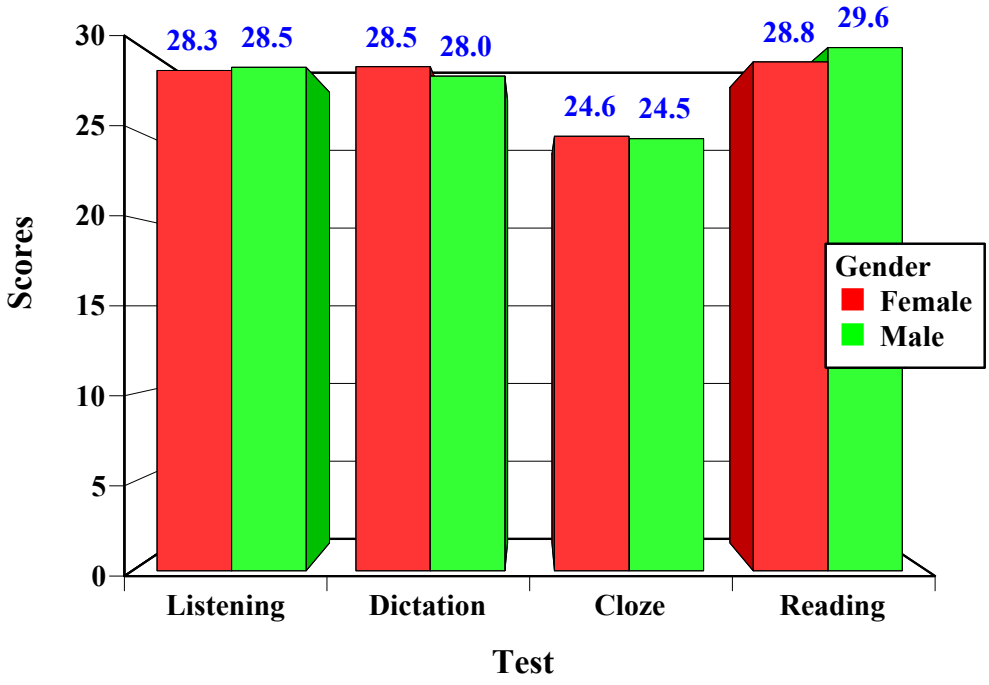


Figure 4 Average scores of examinees by sub-tests and by gender

Bias Analysis of the Sub-Total Scores

Gender

A bar graph of mean scores of every sub-test of each gender is shown in Figure 4. Again, there appears to be no systematic differences and the average scores are fairly closed and almost identical in some sub-tests. It is interesting to observe that male examinees seem to outperform female examinees in reading comprehension test. ANOVA provides confirmation to the visual inspection that no significant difference in group means in every sub-test can be detected.

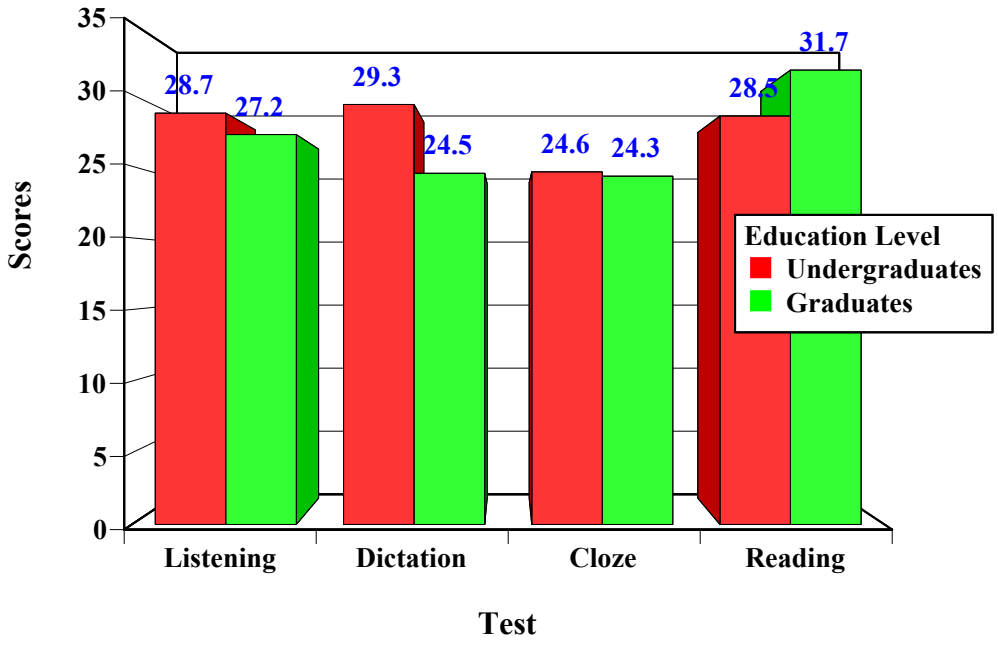


Figure 5 Mean scores of examinees by sub-tests and by education levels

Education Level

It is obvious from the bar graph in Figure 5 that the undergraduate outperform graduate examinees in listening and dictation tests. Particularly, the biggest difference is in dictation where the undergraduates outdo the graduates counterpart by an average of almost five score points. On the contrary, the graduate test takers surpass the undergrads in reading comprehension test by an average of 3.4 score points. The results of the one-way ANOVA

indicate that no significant differences between group means in academic listening and cloze tests. However, the differences in group means of the dictation and reading comprehension tests are statistically significant with F statistics of 6.71 ($d.f.$ = 1/202; $p < 0.025$) for dictation and F statistics of 4.80 ($d.f.$ = 1/203; $p < 0.05$) for reading comprehension test. The differentials in academic listening test and dictation when jointly considered indicates that undergraduate test takers clearly have much better developed listening skills than do the graduates, whereas the graduate examinees have a relatively better developed academic reading ability.

What could possibly explain this seemingly contradictory pattern of performance differentials? Differences in student characteristics may be the answer[6]. Many of the undergraduate examinees are of the Generation 1.5[7], transfer from another university, or freshmen who spent 2-3 years in high school in the US prior to coming to the University of Hawaii. Those students tend to have good English listening ability. On the contrary, their academic writing and reading skills are still in developing stage. The graduate test takers, on the other hand, mostly just arrived from foreign countries. They tend to have limited listening and speaking ability coupled with disproportionately better developed academic reading skills. These characteristics perfectly describe the observed pattern.

[6] Kenton Harsch, the Assistant Director of the ELI, contributes this important insight.
[7] Generation 1.5 refers to students who have mixed characteristics of the first and second generation immigrants (Harklau, 1999). They tend to be immigrant children who arrive in the US at an age before they master their first language. They grow up speaking their mother tongue at home and learning English from social interaction. Whiting (2003) characterizes the Generation 1.5 as those who have no first language. The Gen 1.5 students often appear to be native English speakers in conversation but they may also feel that they have no full command in English (Whiting, 2003).

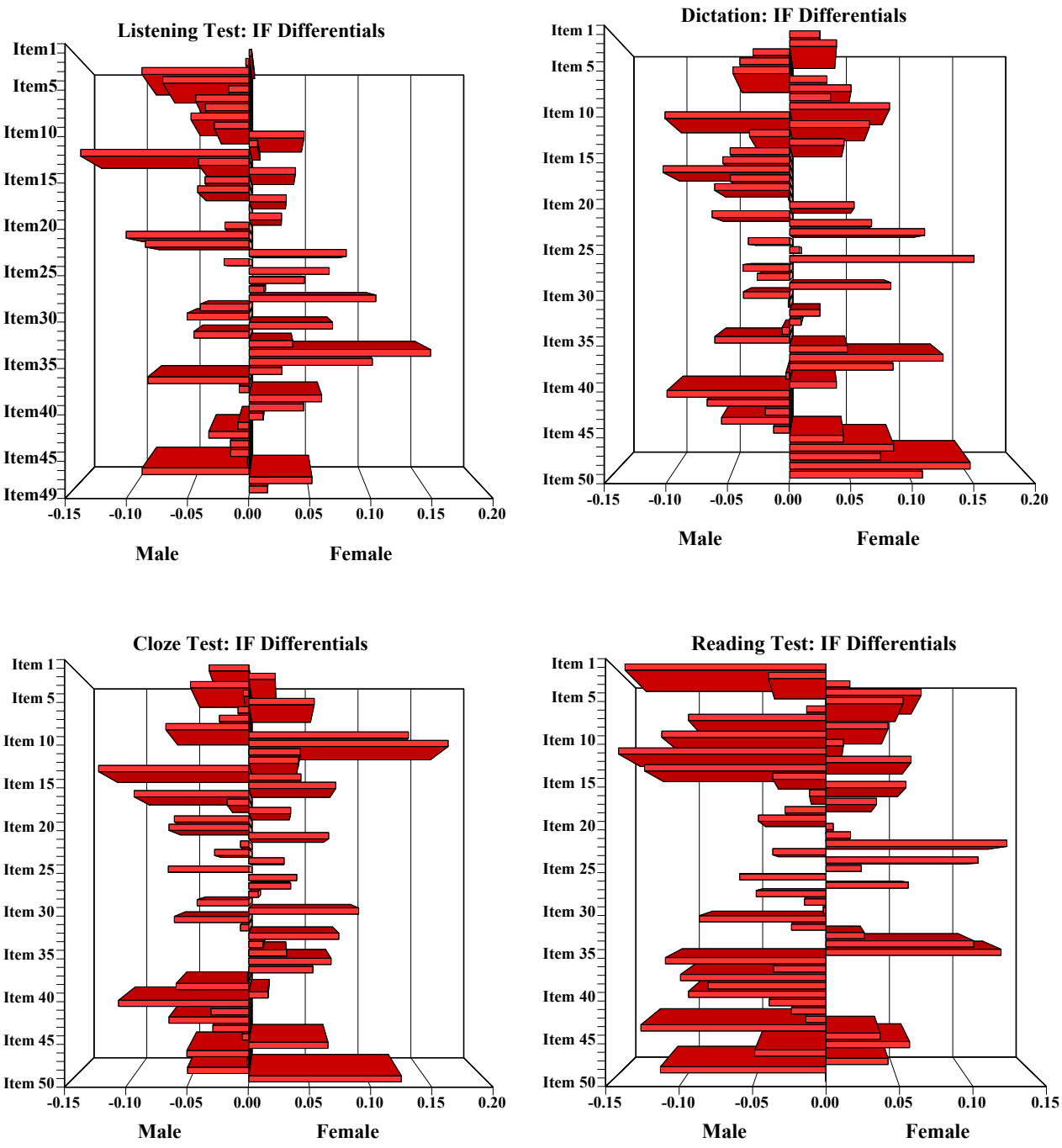


Figure 6 Item Facility (IF) differentials (defined as IF_{Female}-IF_{Male}) by gender and by sub-tests.

DIF Item Analysis

Gender

Figure 6 graphically summarizes IF differentials of every sub-test by gender. The differentials appear to be no systematic patterns. There are only three DIF items as defined by court ruling in the case between Golden Rule Insurance, Co vs. Mathias case, i.e., ID differentials ≥ 0.15. Table 3 reports summary statistics which include *t*-tests that test whether the means of IF differentials are significantly different from zero. The results of tests indicate that gender means of every sub-test do not differ from zero which means that the $IF_{Female} = IF_{Male}$ from statistical perspective. This result is consistent with those obtained by Ryan and Bachman (1992) and Wainer and Lukhele (1997). Both studies examined gender bias in the TOEFL.

Table 3

Summary statistics of IF differentials by gender and education level

Sub-Test	Mean	Min	Max	Between Group Correlation	Shared Variance (r2)
IF diffrentials by gender (IF_{Female}-IF_{Male})					
Academic listening	-0.004	-0.141	0.151	0.958	0.919
Dictation	0.011	-0.105	0.153	0.962	0.926
Cloze	0.003	-0.125	0.166	0.932	0.868
Reading comprehension	-0.016	-0.144	0.126	0.848	0.718
IF differentials by educational enrolment level ($IF_{Undergrad}$-IF_{Grad})					
Academic listenning	0.030*	-0.138	0.227	0.918	0.842
Dictation	0.096*	-0.077	0.279	0.940	0.883
Cloze	0.005	-0.221	0.246	0.844	0.712
Reading comprehension	-0.065*	-0.274	0.093	0.767	0.588

Note: * $p < 0.01$

Education level

Figure 7 depicts performance differentials by education level. There are apparent patterns consistent with the results of ANOVA. While the undergraduates one-sidedly

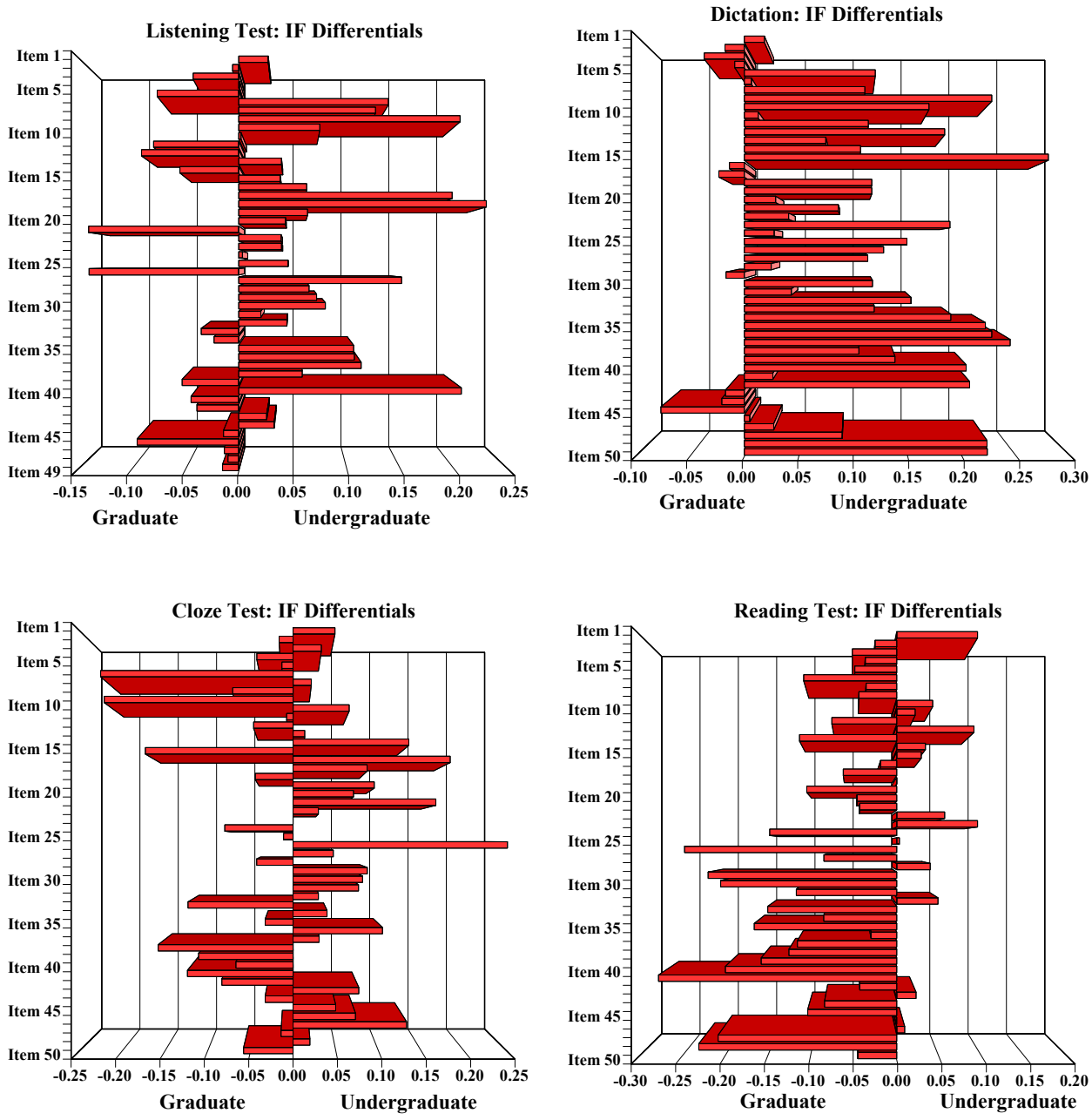


Figure 7 Item Facility (IF) differentials (defined as $IF_{Underdrads}$-IF_{Grads}) by education level and by sub-tests.

outperform the graduates in dictation, the graduates surpass the undergraduates in reading comprehension test by big margins. The *t*-test also indicates that the mean of IF differentials of the academic listening sub-test is statistically different from zero ($p < 0.01$) which means that the $IF_{Undergraduates} > IF_{Graduates}$. This finding is significant because ANOVA did not detect any important difference between undergraduates-graduates in the *scores* of listening test. There are 34 DIF items in various tests with IF differentials of ≥ 0.15. Many of them even meet the EEOC's definition of IF differentials of > 0.20.

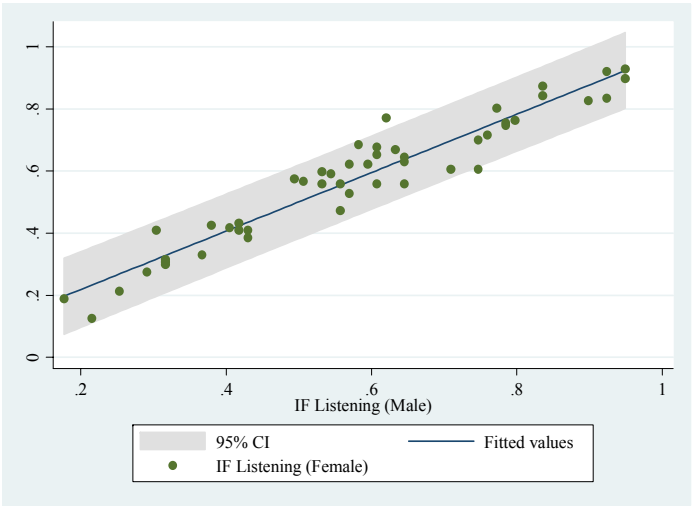
DIF Item Analysis: Outlier detection method

Figure 8 and 9 graphically showed DIF items. Any points outside the 95% confidence interval band are considered outliers and hence DIF items. In all, there are only nine potential DIF items three of which are gender DIF and six of which are education-level DIF. There are much fewer DIF items than those identified by the legal definitions. This outlier detection method is more plausible than is the legal approach. The next crucial step would be reviewing the DIF items to see if they contain language or content bias. This step, however, is outside the scope of this study.

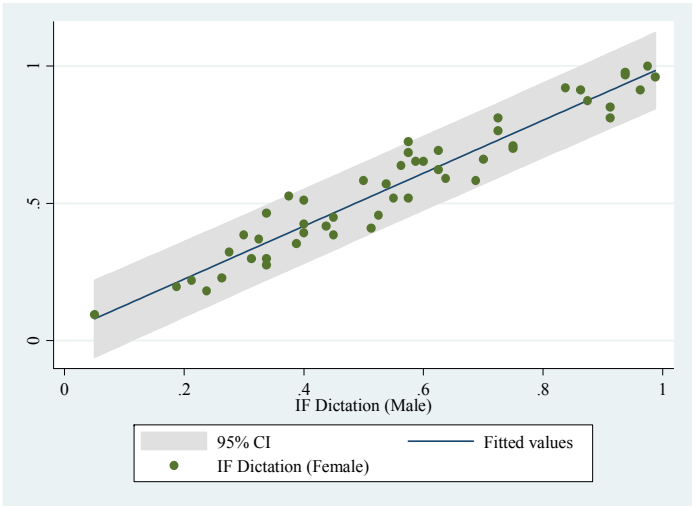
Reliability and Validity of the ELI Placement Test

The notion of reliability rests on the concept of correlation. Figure 10 shows a matrix of a correlograms together with the values of correlation coefficients. For a placement test, the correlation coefficients across sub-tests ranging from 0.518-0.749 are considered low. As expected, the correlation coefficients between sub-tests that are designed to measure the same construct appear to be relatively stronger. For example, correlation coefficient of listening and dictation tests which are meant to test listening ability appears to be the strongest at 0.749.

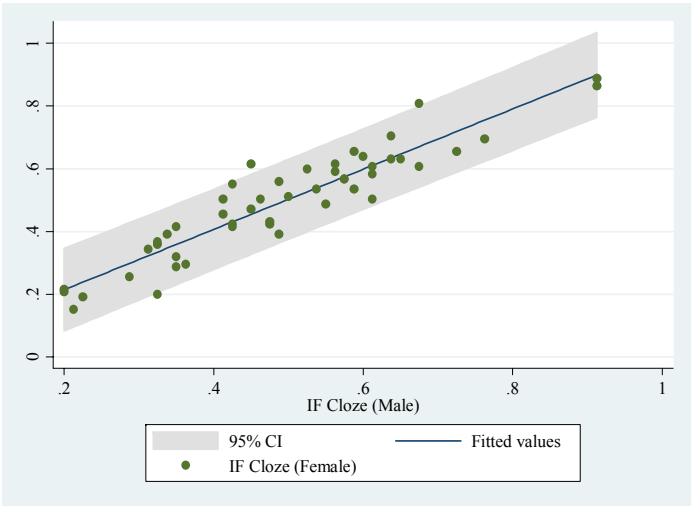
Similarly, the correlation coefficient of cloze and reading comprehension tests which are designed to measure reading ability is the second strongest at 0.670.



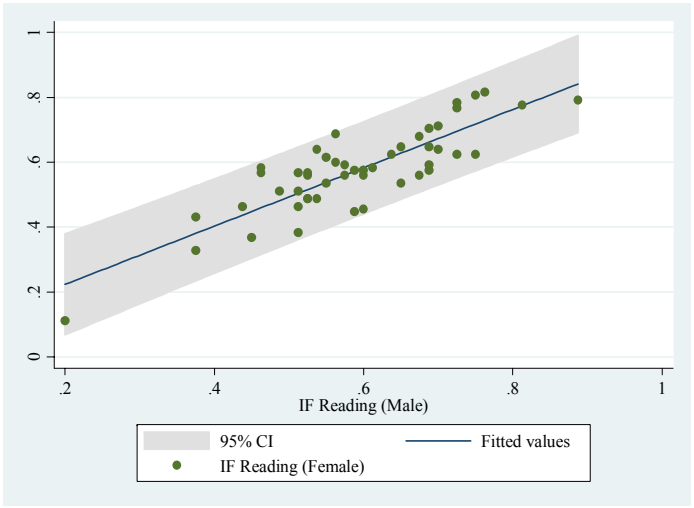
IF of Listening



IF of Dictation

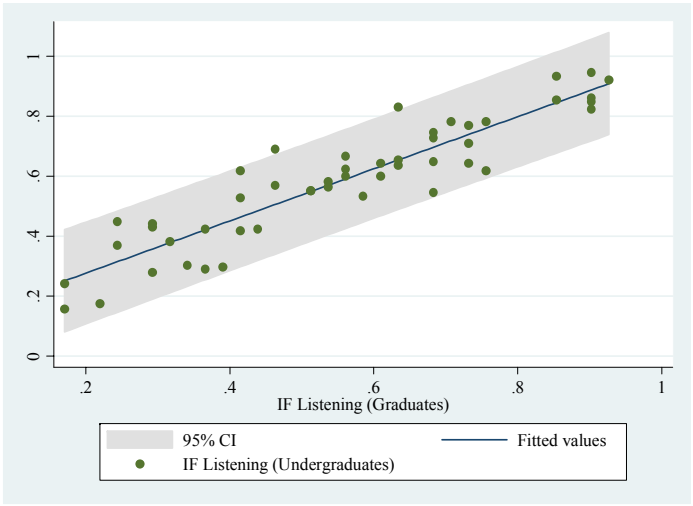


IF of Cloze

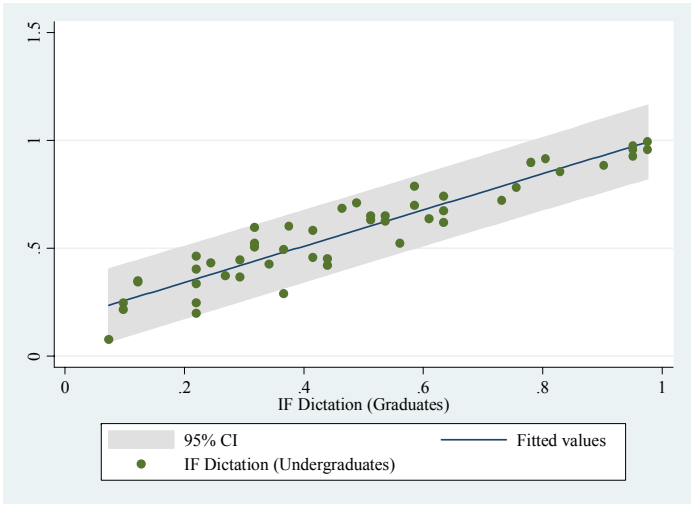


IF of Reading

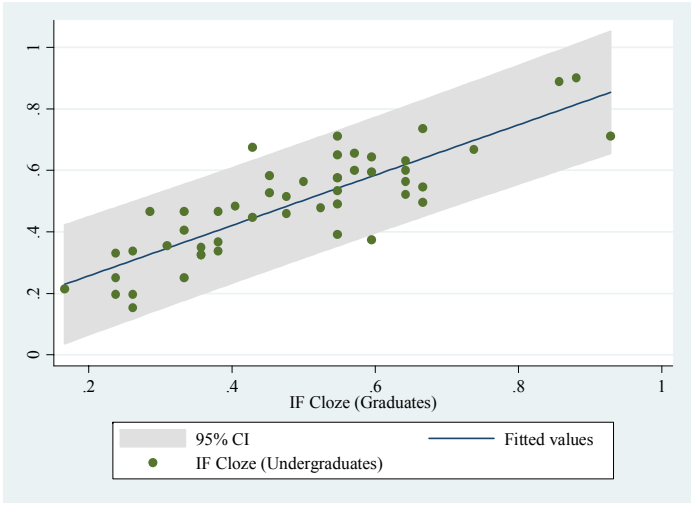
Figure 8 Scatter plot of IF by gender for all sub-tests with linear regression line coupled with 95% confidence interval



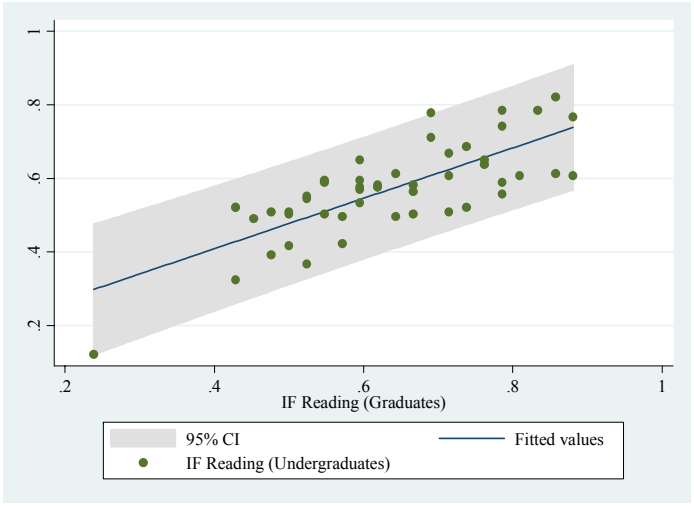
IF of Listening



IF of Dictation



IF of Cloze



IF of Reading

Figure 9 Scatter plot of IF by education level for all sub-tests with linear regression line coupled with 95% confidence interval bands

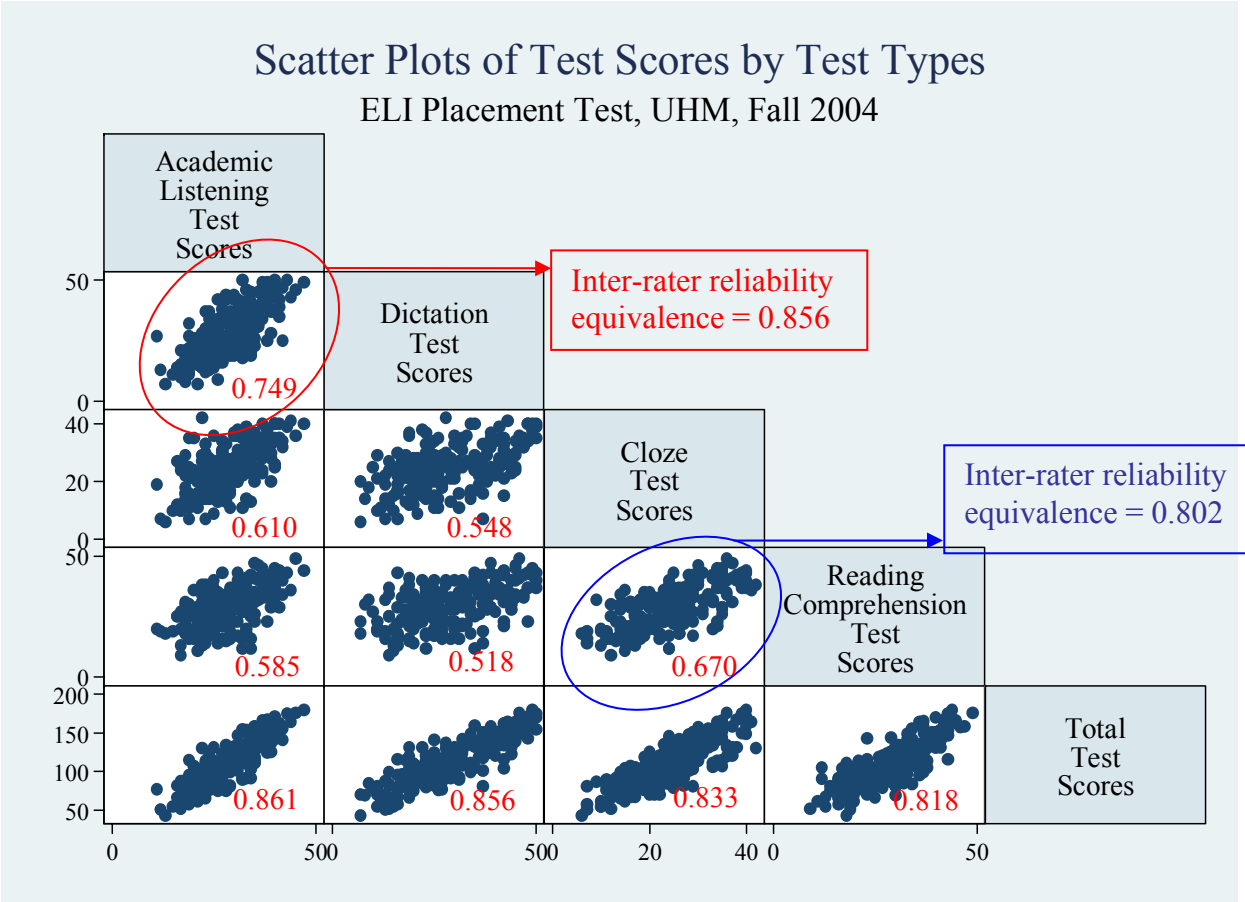


Figure 10 A half matrix of correlograms of total test scores and all sub-tests.

A practice at ELI to allow the higher score among a pair of tests that measure the same construct as a better representation of students' true ability implies that both tests (listening vs. dictation; cloze vs. reading test) are equivalent. The relatively low correlation coefficients mean low shared variance (r^2) and equivalency of the tests as well as their validity are in doubt, especially for cloze and reading tests. Moreover, the low correlation coefficients imply low reliability as evident in the reported the inter-rater reliability estimates of 0.856 and 0.802[8] for listening and reading ability respectively. For a placement test, it is ideal to have reliability in the 0.90s. Table 4 reports internal reliability estimates of each test. All sub-tests but dictation has reliability estimates below 0.90. Overall, however, the ELI placement test exhibits high degree of reliability. The high estimates of all tests combined are primarily the result of having significantly larger number of items.

Table 4

Internal reliability estimates of the ELI placement test, UHM, Fall 2004

Tests	S-B Prophecy	Cronbach α	K-R 20	K-R 21
Cloze	0.8461	0.8465	0.8461	0.8247
Dictation	0.9301	0.9351	0.9352	0.9108
Academic Listening	0.8280	0.8264	0.8226	0.7773
Reading comprehension	0.8634	0.8634	0.8609	0.8505
Reading version A	0.8733	0.8743	0.8713	0.8577
Reading version B	0.8519	0.8523	0.8500	0.8366
Overall	0.9553	0.9556	0.9555	0.9472

Conclusion

This study finds no evidence of gender bias in the ELI placement test. This finding is in line with previous studies in TOEFL [Ryan and Bachman (1992) and Wainer and Lukhele

[8] When there are two or more different test formats but equivalent tests designed to measure the same construct, the tests themselves may be viewed as if they were different raters. Therefore, an equivalence of inter-rater reliability can be estimated from correlation coefficient of the test scores.

(1997)]. Evidences suggest a presence of significant performance differentials between the undergraduate and graduate examinees. However, such differences are likely to closely link to variations in the construct under measured. Therefore, the performance differentials by educational level are *not* likely to be testing bias. The overall reliability of the ELI placement test is respectably high, while, however, its sub-tests' reliability in the 0.80s are slightly less than ideal. Since placement decisions are based on sub-tests rather than on the total scores, it is desirable to examine ways in which consistency of the sub-tests can be raised. More importantly, given that the traditional reading comprehension test is reliable and valid, the low correlation between cloze and reading test and, hence, its inter-rater reliability indicates potential validity problem. This study intends to be exploratory. Its important methodological limitation in identifying DIF items is the absence of control for variations in overall proficiency or ability. Identifying DIF items based on legal definition is arbitrary and unreliable. Evidence based on Kunnan's outlier detection method indicates that only a small numbers of items show potential to be DIF. A few DIF items discovered are not yet a cause for celebrations. Lin and Wu (2003) cited Nandakumar (1993) that items with small or statistically undetectable DIF can be functioning differentially when bias analysis is done at a bundle of items level. In their study of China's academic English Proficiency Test (EPT), Lin and Wu found little evidence of gender item DIF but the gender differential bundle functioning (DBF) was discovered. Future bias study of ELI placement test should examine the DBF.

References

ALTE. (1998). *Multilingual glossary of language testing terms*. Cambridge: Cambridge University Press.

Brown, J. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Davies, A. et al. (1999). *Dictionary of language testing*, Cambridge; New York, NY: Press Syndicate of the University of Cambridge.

Harklau, L. (2003). Generation 1.5 students and college writing. *Digest*. October 2003, from http://www.cal.org/resources/digest/0305harklau.html.

Kunnan, A.J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, 24, 741-746.

Lin, J. & Wu, F. (2003). *Differential performance by gender in foreign language testing*. The Center for Research in Applied Measurement and Evaluation, the University of Alberta.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *16*, 159-176.

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

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement,* 57(5), 741-759

Whiting, S. (2003). The new generation gap: English is hard enough even when it's your first language. *San Francisco Chronicle*. Sunday, December 14, 2003, from http://www.sfgate.com/cgi-bin/article.cgi?file=/chronicle/archive/2003/12/14/CM199288.DTL.

Appendix

Table A1

Item facility (IF) differentials of listening and dictation tests by gender; ELI Fall, 2004

Test Item	IF Listening			IF Dictation		
	IF Female	IF Male	IF_{Female}-IF_{Male}	IF Female	IF Male	IF_{Female}-IF_{Male}
Item1	0.559	0.557	0.002	1.000	0.975	0.025
Item2	0.921	0.924	-0.003	0.976	0.938	0.039
Item3	0.835	0.924	-0.089	0.520	0.550	-0.030
Item4	0.827	0.899	-0.072	0.709	0.750	-0.041
Item5	0.299	0.316	-0.017	0.591	0.638	-0.047
Item6	0.386	0.430	-0.045	0.969	0.938	0.031
Item7	0.331	0.367	-0.036	0.913	0.863	0.051
Item8	0.559	0.608	-0.049	0.571	0.538	0.034
Item9	0.756	0.785	-0.029	0.583	0.500	0.083
Item10	0.654	0.608	0.046	0.409	0.513	-0.103
Item11	0.843	0.835	0.007	0.654	0.588	0.066
Item12	0.606	0.747	-0.141	0.354	0.388	-0.033
Item13	0.528	0.570	-0.042	0.370	0.325	0.045
Item14	0.874	0.835	0.039	0.701	0.750	-0.049
Item15	0.748	0.785	-0.037	0.520	0.575	-0.055
Item16	0.717	0.759	-0.043	0.583	0.688	-0.105
Item17	0.803	0.772	0.031	0.913	0.963	-0.049
Item18	0.646	0.646	0.000	0.850	0.913	-0.062
Item19	0.622	0.595	0.027	0.874	0.875	-0.001
Item20	0.929	0.949	-0.020	0.654	0.600	0.054
Item21	0.606	0.709	-0.103	0.386	0.450	-0.064
Item22	0.559	0.646	-0.087	0.693	0.625	0.068
Item23	0.575	0.494	0.081	0.512	0.400	0.112
Item24	0.409	0.430	-0.021	0.228	0.263	-0.034
Item25	0.598	0.532	0.067	0.220	0.213	0.008
Item26	0.591	0.544	0.046	0.528	0.375	0.153
Item27	0.417	0.405	0.012	0.661	0.700	-0.039
Item28	0.409	0.304	0.106	0.961	0.988	-0.027
Item29	0.213	0.253	-0.041	0.921	0.838	0.084
Item30	0.898	0.949	-0.052	0.299	0.338	-0.038
Item31	0.677	0.608	0.070	0.449	0.450	-0.001
Item32	0.701	0.747	-0.046	0.425	0.400	0.025
Item33	0.669	0.633	0.036	0.197	0.188	0.009
Item34	0.772	0.620	0.151	0.394	0.400	-0.006
Item35	0.685	0.582	0.103	0.276	0.338	-0.062

Test Item	IF Listening			IF Dictation		
	IF Female	IF Male	IF_{Female}-IF_{Male}	IF Female	IF Male	IF_{Female}-IF_{Male}
Item36	0.559	0.532	0.027	0.323	0.275	0.048
Item37	0.472	0.557	-0.085	0.465	0.338	0.127
Item38	0.409	0.418	-0.008	0.386	0.300	0.086
Item39	0.567	0.506	0.061	0.622	0.625	-0.003
Item40	0.425	0.380	0.045	0.764	0.725	0.039
Item41	0.189	0.177	0.012	0.811	0.913	-0.101
Item42	0.307	0.316	-0.009	0.457	0.525	-0.068
Item43	0.764	0.797	-0.034	0.417	0.438	-0.020
Item44	0.630	0.646	-0.016	0.181	0.238	-0.056
Item45	0.276	0.291	-0.016	0.299	0.313	-0.013
Item46	0.315	0.316	-0.001	0.094	0.050	0.044
Item47	0.126	0.215	-0.089	0.811	0.725	0.086
Item48	0.622	0.570	0.052	0.638	0.563	0.075
Item49	0.433	0.418	0.015	0.724	0.575	0.149
Item 50				0.685	0.575	0.110
			Descriptive Statistics			
Mean	0.578	0.582	-0.004	0.571	0.560	0.011
Minimum	0.126	0.177	-0.141	0.094	0.050	-0.105
Maximum	0.929	0.949	0.151	1.000	0.988	0.153

Table A2

Item facility (IF) differentials of cloze and reading comprehension tests by gender; ELI Fall, 2004

Test Item	IF Cloze Test			IF Reading Comprehension Test		
	IF Female	IF Male	IF_{Female}-IF_{Male}	IF Female	IF Male	IF_{Female}-IF_{Male}
Item 1	0.192	0.225	-0.033	0.448	0.588	-0.140
Item 2	0.472	0.450	0.022	0.560	0.600	-0.040
Item 3	0.864	0.913	-0.049	0.704	0.688	0.017
Item 4	0.608	0.613	-0.005	0.616	0.550	0.066
Item 5	0.392	0.338	0.055	0.816	0.763	0.054
Item 6	0.416	0.425	-0.009	0.624	0.638	-0.014
Item 7	0.888	0.913	-0.025	0.792	0.888	-0.095
Item 8	0.656	0.725	-0.069	0.768	0.725	0.043
Item 9	0.808	0.675	0.133	0.536	0.650	-0.114
Item 10	0.616	0.450	0.166	0.712	0.700	0.012
Item 11	0.368	0.325	0.043	0.456	0.600	-0.144
Item 12	0.504	0.463	0.042	0.784	0.725	0.059
Item 13	0.200	0.325	-0.125	0.624	0.750	-0.126
Item 14	0.456	0.413	0.044	0.488	0.525	-0.037

Test Item	IF Cloze Test			IF Reading Comprehension Test		
	IF Female	IF Male	IF_{Female}-IF_{Male}	IF Female	IF Male	IF_{Female}-IF_{Male}
Item 15	0.560	0.488	0.073	0.568	0.513	0.056
Item 16	0.392	0.488	-0.096	0.576	0.588	-0.012
Item 17	0.632	0.650	-0.018	0.560	0.525	0.035
Item 18	0.360	0.325	0.035	0.584	0.613	-0.029
Item 19	0.288	0.350	-0.062	0.328	0.375	-0.047
Item 20	0.696	0.763	-0.067	0.680	0.675	0.005
Item 21	0.704	0.638	0.067	0.592	0.575	0.017
Item 22	0.568	0.575	-0.007	0.688	0.563	0.126
Item 23	0.584	0.613	-0.029	0.488	0.525	-0.037
Item 24	0.592	0.563	0.030	0.568	0.463	0.106
Item 25	0.608	0.675	-0.067	0.512	0.488	0.025
Item 26	0.640	0.600	0.040	0.640	0.700	-0.060
Item 27	0.360	0.325	0.035	0.432	0.375	0.057
Item 28	0.208	0.200	0.008	0.464	0.513	-0.048
Item 29	0.432	0.475	-0.043	0.560	0.575	-0.015
Item 30	0.504	0.413	0.092	0.648	0.650	-0.002
Item 31	0.488	0.550	-0.062	0.112	0.200	-0.088
Item 32	0.568	0.575	-0.007	0.576	0.600	-0.024
Item 33	0.600	0.525	0.075	0.464	0.438	0.027
Item 34	0.512	0.500	0.012	0.640	0.538	0.103
Item 35	0.344	0.313	0.032	0.584	0.463	0.122
Item 36	0.656	0.588	0.069	0.576	0.688	-0.112
Item 37	0.616	0.563	0.054	0.776	0.813	-0.037
Item 38	0.424	0.425	-0.001	0.624	0.725	-0.101
Item 39	0.152	0.213	-0.061	0.368	0.450	-0.082
Item 40	0.216	0.200	0.016	0.592	0.688	-0.096
Item 41	0.504	0.613	-0.109	0.648	0.688	-0.040
Item 42	0.256	0.288	-0.032	0.576	0.600	-0.024
Item 43	0.296	0.363	-0.067	0.536	0.550	-0.014
Item 44	0.320	0.350	-0.030	0.384	0.513	-0.129
Item 45	0.632	0.638	-0.005	0.600	0.563	0.038
Item 46	0.416	0.350	0.066	0.808	0.750	0.058
Item 47	0.536	0.588	-0.052	0.488	0.538	-0.050
Item 48	0.536	0.538	-0.001	0.568	0.525	0.043
Item 49	0.424	0.475	-0.051	0.560	0.675	-0.115
Item 50	0.552	0.425	0.127	0.512	0.513	0.000
			Descriptive Statistics			
Mean	0.492	0.489	0.003	0.576	0.592	-0.016
Minimum	0.152	0.200	-0.125	0.112	0.200	-0.144
Maximum	0.888	0.913	0.166	0.816	0.888	0.126

Table A3

Item facility (IF) differentials of listening and dictation tests by educational enrollment; ELI Fall, 2004

Test Item	IF Listening			IF Dictation		
	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}
Item 1	0.564	0.537	0.027	0.994	0.976	0.018
Item 2	0.921	0.927	-0.006	0.958	0.976	-0.018
Item 3	0.861	0.902	-0.042	0.524	0.561	-0.037
Item 4	0.855	0.854	0.001	0.723	0.732	-0.009
Item 5	0.291	0.366	-0.075	0.633	0.512	0.120
Item 6	0.430	0.293	0.138	0.958	0.951	0.007
Item 7	0.370	0.244	0.126	0.916	0.805	0.111
Item 8	0.618	0.415	0.204	0.602	0.375	0.227
Item 9	0.782	0.707	0.075	0.584	0.415	0.170
Item 10	0.636	0.634	0.002	0.452	0.439	0.013
Item 11	0.824	0.902	-0.078	0.651	0.537	0.114
Item 12	0.642	0.732	-0.089	0.404	0.220	0.184
Item 13	0.552	0.512	0.039	0.367	0.293	0.075
Item 14	0.848	0.902	-0.054	0.741	0.634	0.107
Item 15	0.770	0.732	0.038	0.596	0.317	0.279
Item 16	0.745	0.683	0.063	0.620	0.634	-0.014
Item 17	0.830	0.634	0.196	0.928	0.951	-0.024
Item 18	0.691	0.463	0.227	0.898	0.780	0.117
Item 19	0.624	0.561	0.063	0.898	0.780	0.117
Item 20	0.945	0.902	0.043	0.639	0.610	0.029
Item 21	0.618	0.756	-0.138	0.428	0.341	0.086
Item 22	0.600	0.561	0.039	0.675	0.634	0.041
Item 23	0.552	0.512	0.039	0.506	0.317	0.189
Item 24	0.418	0.415	0.004	0.247	0.220	0.027
Item 25	0.582	0.537	0.045	0.247	0.098	0.149
Item 26	0.545	0.683	-0.137	0.494	0.366	0.128
Item 27	0.442	0.293	0.150	0.699	0.585	0.113
Item 28	0.382	0.317	0.065	0.976	0.951	0.025
Item 29	0.242	0.171	0.072	0.886	0.902	-0.017
Item 30	0.933	0.854	0.080	0.337	0.220	0.118
Item 31	0.655	0.634	0.020	0.458	0.415	0.043
Item 32	0.727	0.683	0.044	0.446	0.293	0.153
Item 33	0.648	0.683	-0.034	0.217	0.098	0.119
Item 34	0.709	0.732	-0.023	0.434	0.244	0.190
Item 35	0.667	0.561	0.106	0.343	0.122	0.221
Item 36	0.570	0.463	0.106	0.349	0.122	0.227

Test Item	IF Listening			IF Dictation		
	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}
Item 37	0.527	0.415	0.113	0.464	0.220	0.244
Item 38	0.424	0.366	0.058	0.373	0.268	0.105
Item 39	0.533	0.585	-0.052	0.651	0.512	0.138
Item 40	0.448	0.244	0.205	0.789	0.585	0.204
Item 41	0.176	0.220	-0.044	0.855	0.829	0.026
Item 42	0.303	0.341	-0.038	0.524	0.317	0.207
Item 43	0.782	0.756	0.026	0.422	0.439	-0.017
Item 44	0.642	0.610	0.033	0.199	0.220	-0.021
Item 45	0.279	0.293	-0.014	0.289	0.366	-0.077
Item 46	0.297	0.390	-0.093	0.078	0.073	0.005
Item 47	0.158	0.171	-0.013	0.783	0.756	0.027
Item 48	0.600	0.610	-0.010	0.627	0.537	0.090
Item 49	0.424	0.439	-0.015	0.711	0.488	0.223
Item 50				0.687	0.463	0.223
			Descriptive Statistics			
Mean	0.586	0.555	0.030	0.586	0.490	0.096
Minimum	0.158	0.171	-0.138	0.078	0.073	-0.077
Maximum	0.945	0.927	0.227	0.994	0.976	0.279

Table A4

Item facility (IF) differentials of cloze and reading comprehension tests by education enrollment; ELI Fall, 2004

Test Item	IF Cloze Test			IF Reading Comprehension Test		
	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}
Item 1	0.215	0.167	0.048	0.521	0.429	0.093
Item 2	0.460	0.476	-0.016	0.571	0.595	-0.025
Item 3	0.890	0.857	0.032	0.687	0.738	-0.051
Item 4	0.601	0.643	-0.042	0.583	0.619	-0.036
Item 5	0.368	0.381	-0.013	0.785	0.833	-0.048
Item 6	0.374	0.595	-0.221	0.607	0.714	-0.107
Item 7	0.902	0.881	0.021	0.822	0.857	-0.035
Item 8	0.669	0.738	-0.069	0.742	0.786	-0.043
Item 9	0.712	0.929	-0.217	0.589	0.548	0.041
Item 10	0.564	0.500	0.064	0.712	0.690	0.021
Item 11	0.350	0.357	-0.007	0.497	0.571	-0.074
Item 12	0.479	0.524	-0.045	0.779	0.690	0.089

Test Item	IF Cloze Test			IF Reading Comprehension Test		
	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}
Item 13	0.252	0.238	0.013	0.650	0.762	-0.112
Item 14	0.466	0.333	0.133	0.509	0.476	0.033
Item 15	0.497	0.667	-0.170	0.552	0.524	0.028
Item 16	0.466	0.286	0.181	0.577	0.595	-0.019
Item 17	0.656	0.571	0.085	0.534	0.595	-0.061
Item 18	0.337	0.381	-0.044	0.595	0.595	0.000
Item 19	0.331	0.238	0.093	0.325	0.429	-0.103
Item 20	0.736	0.667	0.070	0.669	0.714	-0.046
Item 21	0.712	0.548	0.164	0.577	0.619	-0.042
Item 22	0.577	0.548	0.029	0.650	0.595	0.055
Item 23	0.595	0.595	0.000	0.521	0.429	0.093
Item 24	0.564	0.643	-0.078	0.497	0.643	-0.146
Item 25	0.632	0.643	-0.011	0.503	0.500	0.003
Item 26	0.675	0.429	0.246	0.613	0.857	-0.244
Item 27	0.356	0.310	0.046	0.393	0.476	-0.084
Item 28	0.196	0.238	-0.042	0.491	0.452	0.038
Item 29	0.466	0.381	0.085	0.521	0.738	-0.217
Item 30	0.485	0.405	0.080	0.607	0.810	-0.202
Item 31	0.528	0.452	0.075	0.123	0.238	-0.115
Item 32	0.577	0.548	0.029	0.595	0.548	0.047
Item 33	0.546	0.667	-0.121	0.423	0.571	-0.148
Item 34	0.515	0.476	0.039	0.583	0.667	-0.084
Item 35	0.325	0.357	-0.032	0.503	0.667	-0.164
Item 36	0.650	0.548	0.103	0.613	0.643	-0.029
Item 37	0.601	0.571	0.030	0.767	0.881	-0.114
Item 38	0.393	0.548	-0.155	0.638	0.762	-0.124
Item 39	0.153	0.262	-0.109	0.368	0.524	-0.156
Item 40	0.196	0.262	-0.066	0.589	0.786	-0.197
Item 41	0.521	0.643	-0.121	0.607	0.881	-0.274
Item 42	0.252	0.333	-0.082	0.577	0.619	-0.042
Item 43	0.337	0.262	0.076	0.546	0.524	0.022
Item 44	0.325	0.357	-0.032	0.417	0.500	-0.083
Item 45	0.644	0.595	0.049	0.564	0.667	-0.102
Item 46	0.405	0.333	0.072	0.785	0.786	0.000
Item 47	0.583	0.452	0.130	0.509	0.500	0.009
Item 48	0.534	0.548	-0.014	0.509	0.714	-0.205
Item 49	0.448	0.429	0.019	0.558	0.786	-0.227
Item 50	0.491	0.548	-0.057	0.503	0.548	-0.045
			Descriptive Statistics			
Mean	0.492	0.487	0.005	0.569	0.634	-0.065

Test Item	IF Cloze Test			IF Reading Comprehension Test		
	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}	IF Undergrad	IF Grad	$IF_{Undergrad}$-IF_{Grad}
Minimum	0.153	0.167	-0.221	0.123	0.238	-0.274
Maximum	0.902	0.929	0.246	0.822	0.881	0.093

Identifying DIF Items in Gender

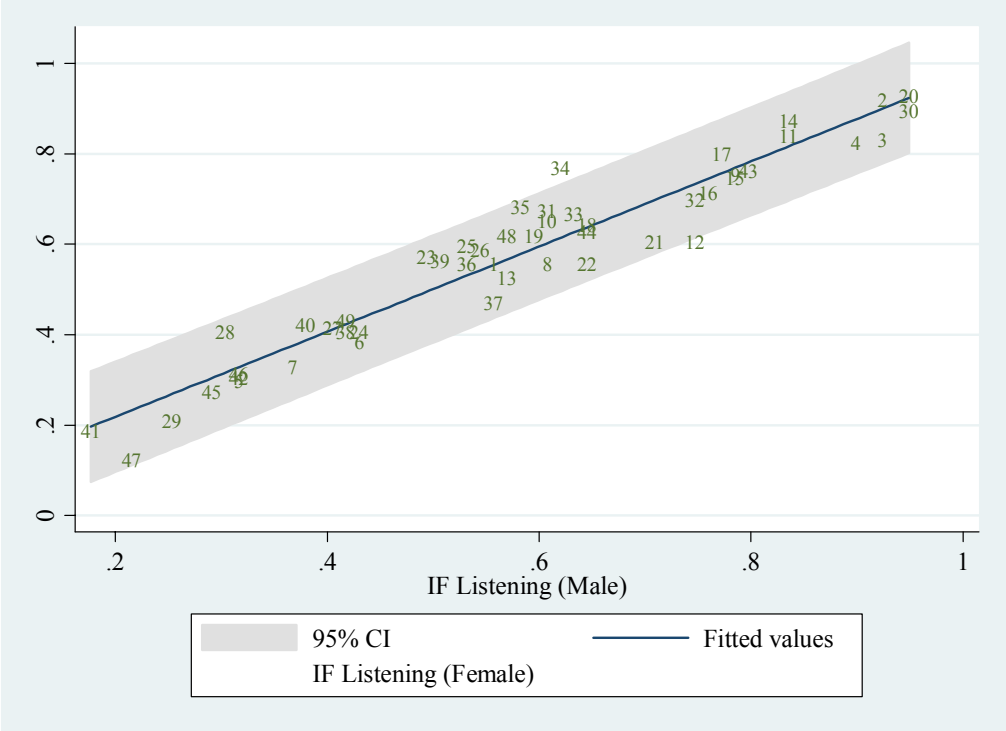


Figure 1A: Identifying DIF item in academic listening test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

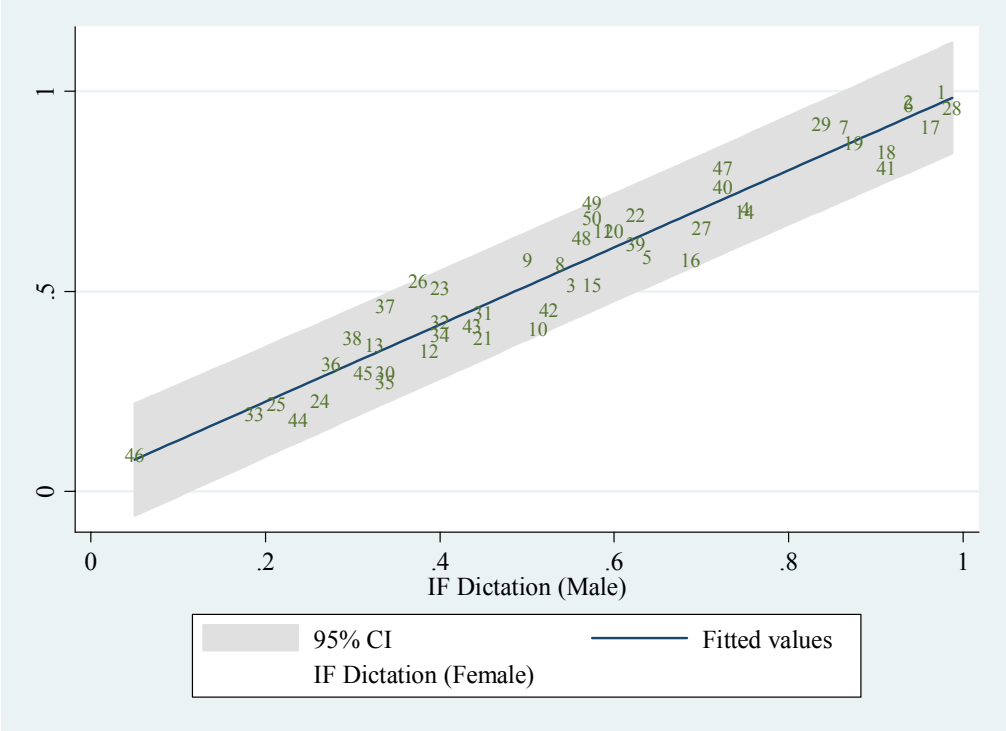


Figure 2A: Identifying DIF item in dictation test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

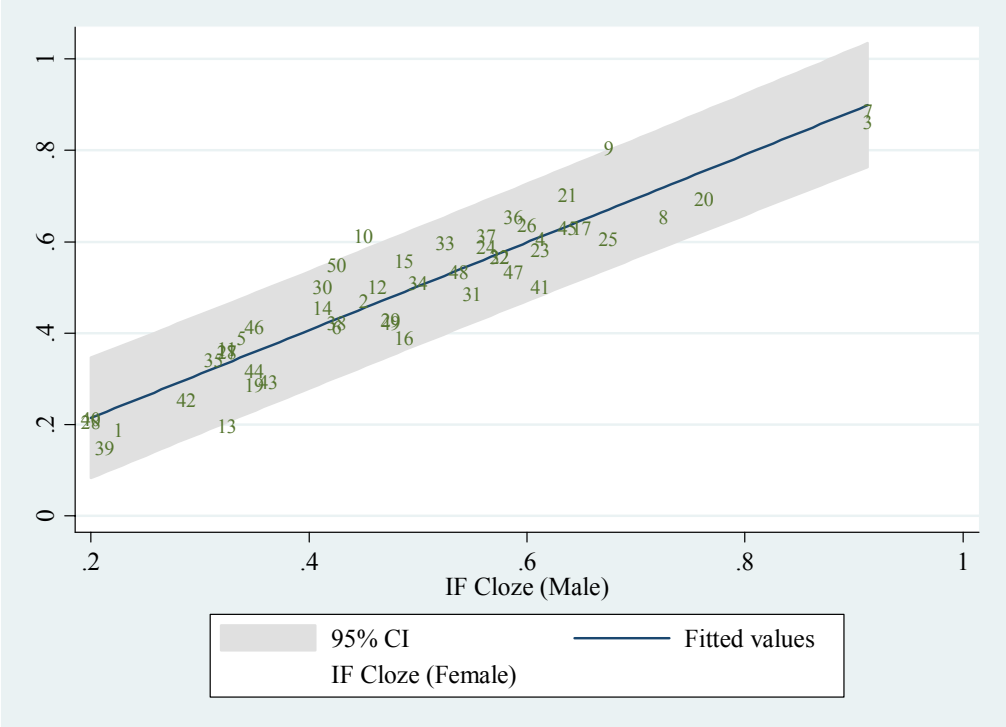


Figure 3A: Identifying DIF item in cloze test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

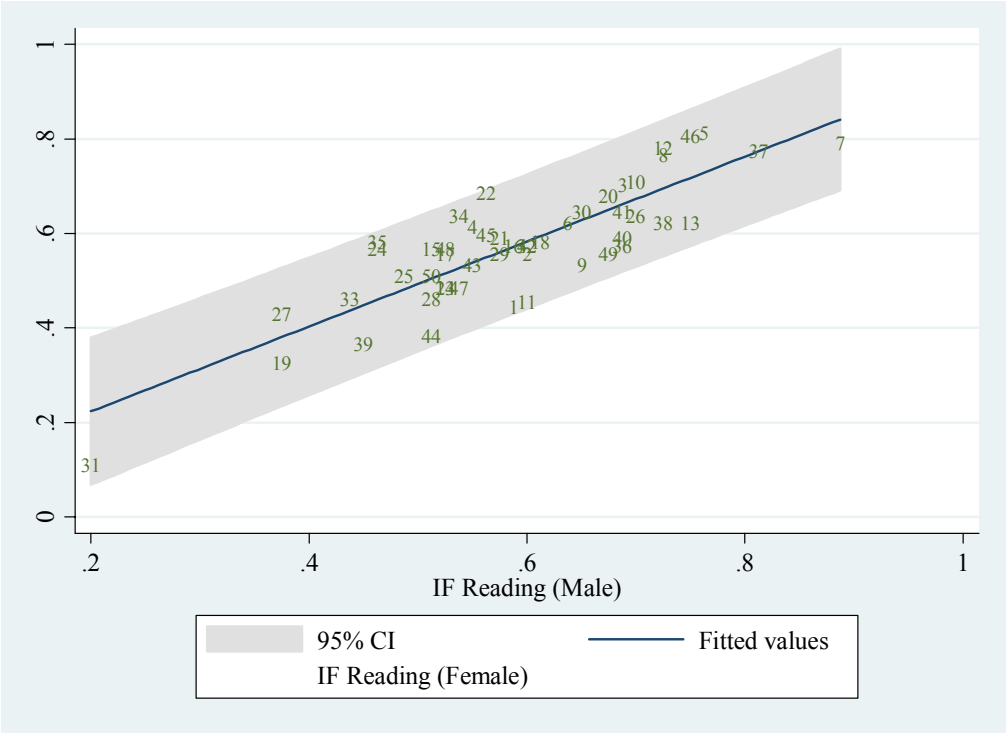


Figure 4A: Identifying DIF item in cloze test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

Identifying DIF Items in Undergraduate-Graduate Enrollment Level

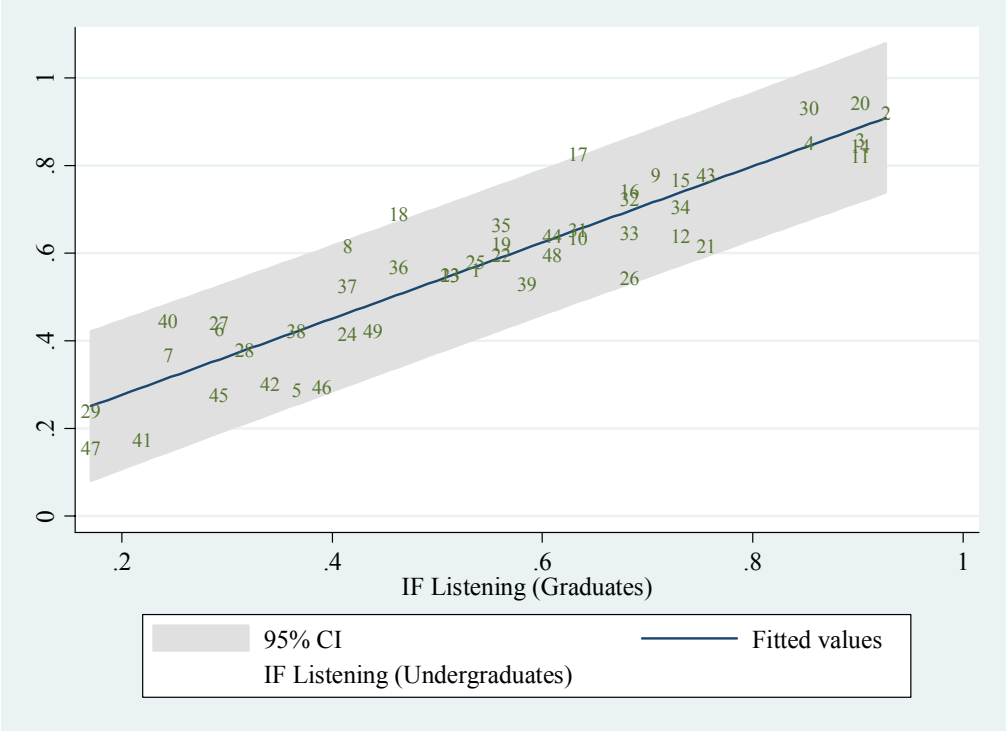


Figure 5A: Identifying DIF item in listening test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

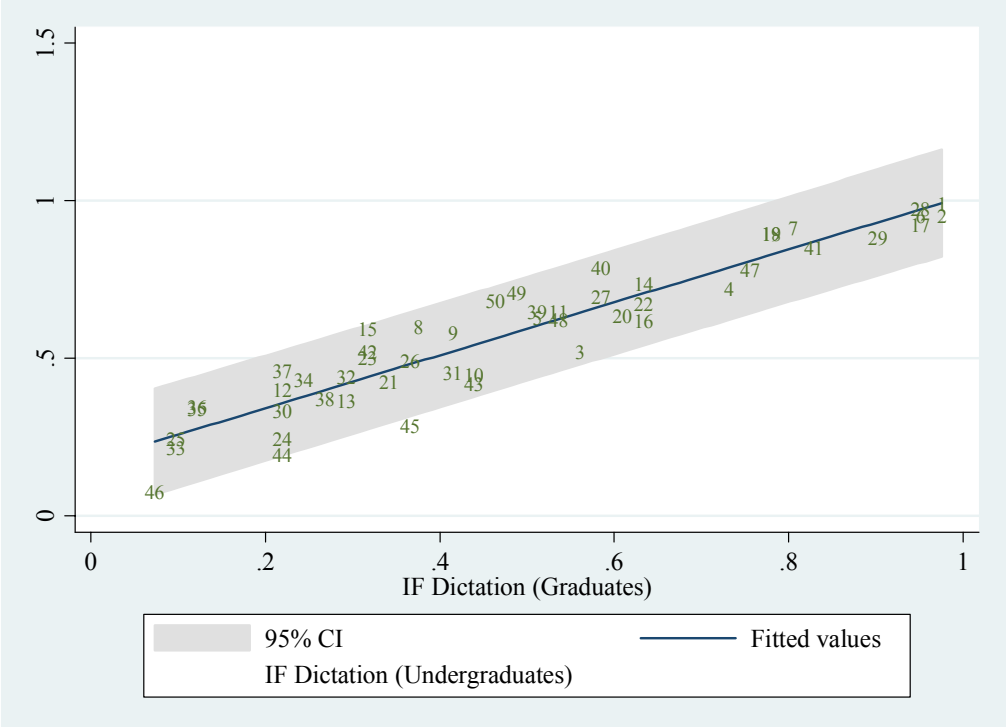


Figure 6A: Identifying DIF item in dictation test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

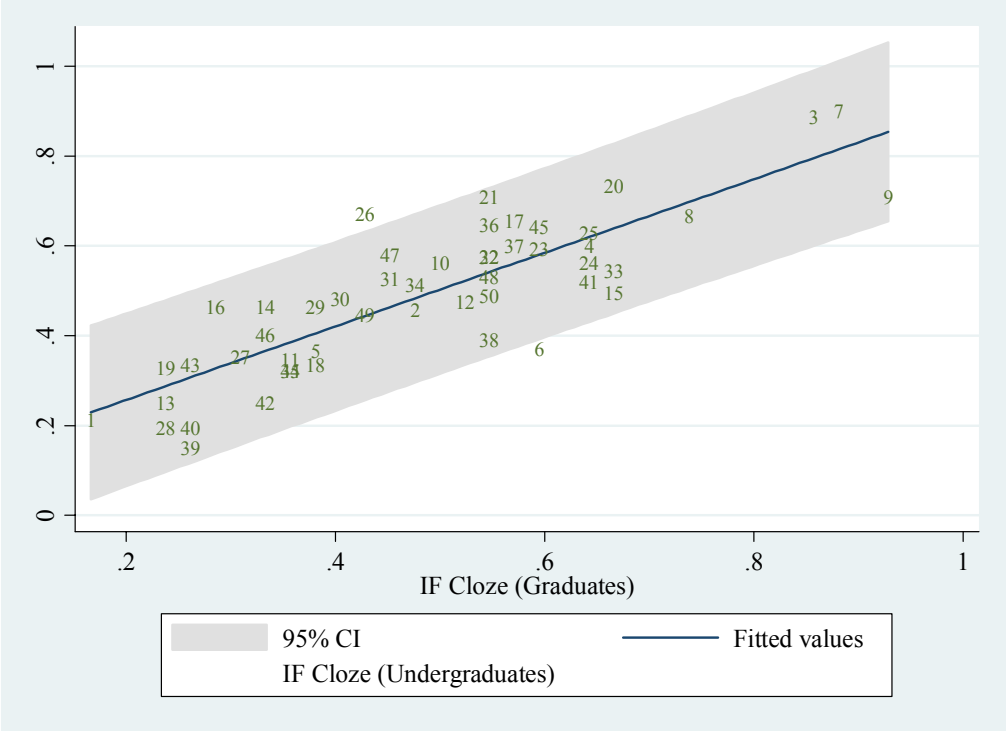


Figure 7A: Identifying DIF item in cloze test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.

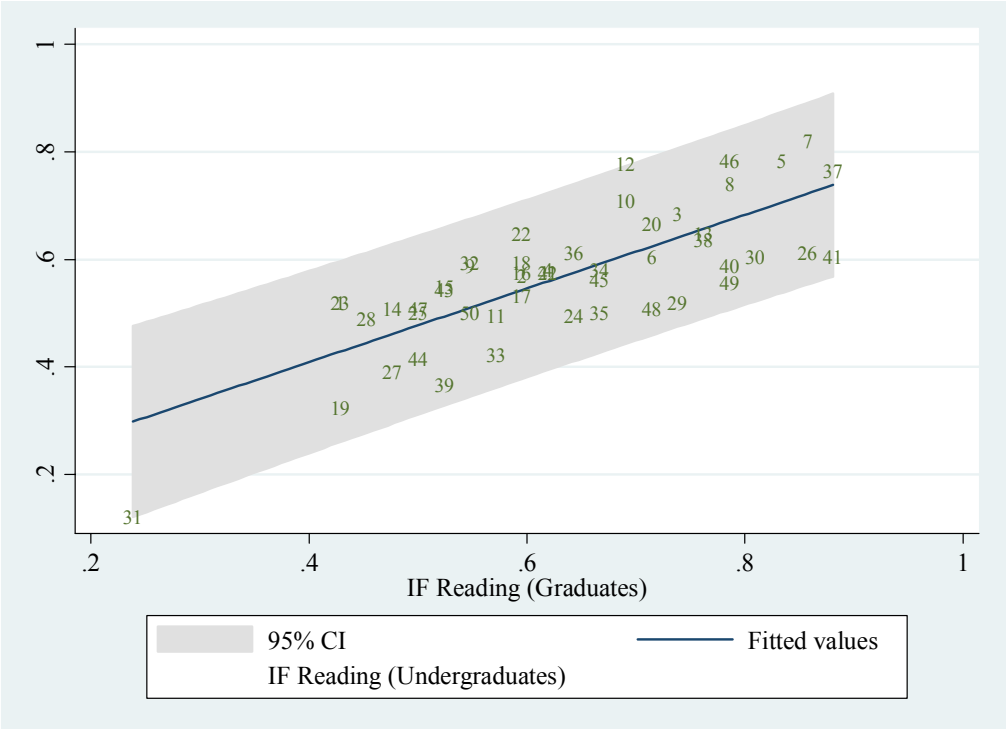


Figure 8A: Identifying DIF item in cloze test using outlier detection method. Observations outside 95% confidence interval band are considered DIF item.