Assessing Unidimensionality and Differential Item Functioning in Qualifying Examination for Senior Secondary School Students, Osun State, Nigeria

Taiwo Oluwafemi Ajeigbe[1,*] & Eyitayo Rufus Ifedayo Afolabi[1]

[1]Department of Education Foundations and Counselling, Faculty of Education, Obafemi Awolowo Univerisity, Ile-Ife, Osun State, Nigeria

*Corresponding author: Department of Education Foundations and Counselling, Faculty of Education, Obafemi Awolowo Univerisity, Ile-Ife, Osun State, Nigeria. E-mail: taiaje@oauife.edu.ng

Received: May 14, 2014 Accepted: June 16, 2014 Online Published: June 29, 2014

doi:10.5430/wje.v4n4p30 URL: http://dx.doi.org/10.5430/wje.v4n4p30

Abstract

This study assessed unidimensionality and occurrence of Differential Item Functioning (DIF) in Mathematics and English Language items of Osun State Qualifying Examination. The study made use of secondary data. The results showed that OSQ Mathematics ($-0.094 \leq r \leq 0.236$) and English Language items ($-0.095 \leq r \leq 0.228$) were unidimensional. Also, there was occurrence of DIF items in both Mathematics and English Language multiple-choice items of the OSQE for 2008. Fourteen items representing 28% of the 50 items in the Mathematics examination exhibited DIF and 10 items, representing 20% of the 50 items in the English Language examination exhibited DIF. The study concluded that the Examination contained considerable number of items that exhibited DIF and therefore requires adequate item quality improvement to justify its use as the inclusion or exclusion criterion of state candidate in West African Examination Council.

Keywords: *unidimensionality; differential item functioning; item quality*

1. Introduction

In education, students' success is established through tests or examinations. The examination could be used for promotion, recruitment, placement and so on via valid and reliable instrument of measurement (items). Test experts are expected to generate good test items that could be used to examine the ability of test-takers from whether homogenous or heterogeneous settings, as the value of such measure would be domiciled in its quality. To ensure quality of test items, such items should not behave differently for particular subgroups of test-takers. If an item functions differently for certain groups, the item reduces the validity of the measure for that construct, and test fairness is threatened. In Rasch measurement model, test items which are biased toward different subgroups within a given populationas a results of unintended factors, such as ability, gender, and ethnicity, subgroups, will exhibit Differential Item Functioning (DIF).

The assumptions of the Rasch model include unidimensionality (i.e., whether the items form a unitary latent trait) and local independence (i.e., the likelihood of the person correctly answering to an item is independent from the other items in the test; Green, 1996; Lee, 1997). Unidimensionality and local independence are assessed using fit statistics, which report the extent to which the pattern of observed answers and the modeled expectations are evaluated in terms of item fit and person fit to the Rasch model (Sick, 2010).Unidimensionality occurs when each of the items in a test measures a single trait, which in principle assumed that local independence. Local independence is achieved when testees' abilitiesto responses to items is independent of one another. This means that ability to respond correctly on an item is influenced by any other item(s) in the test.

The study of DIF has become an integral part of determining the validity and reliability of standardized tests. In the context of tests, DIF occurs when people from different groups with the same ability have systematically different responses to specific test items. If, for example, in a mathematics test, boys display higher probability of answering correctly more often than girls of equal ability level because the contents in the test items are biased against girls, then

the items are said to exhibit DIF and should be considered for modification or removal from the test. Differential item functioning of an item can therefore be understood as a lack of conditional independence between an item response and group membership (often gender, location or ethnicity) given the same latent ability or trait.

When standardized tests are administered on test takers, the test-taking population could vary on a number of personal and educational characteristics such as age, gender, first language, environment, and academic discipline. From the researcher's personal experience and observations, some test developers do not always take into cognizance the diversities that characterized the test takers before administering such test. This could result into various kinds of errors especially scoring error that inflates scores for one group at the expense of the other. Consequently, such test may be regarded as unreliable or lack test fairness. In a standardized test, item characteristics such as difficulty index, discrimination index, reliability, and validity must have been determined or established before they are administered on students. However, in this study, the researcher is of the opinion that the Osun State Qualifying Examination (OSQE) which may be assumed to be a standardized examination should be examined to ascertain the extent to which the test items are DIF free, bias free and fair.

In this study, the state examination that constituted area of focus was Osun State Qualifying Examination (OSQE) which was introduced in 2004 as an intervention measure to arrest the decline and enhance better performance of students in public examinations. The Osun State Government instituted qualifying examination for SS II students in public secondary schools. It is only those students that pass the State Qualifying Examination (SQE) that will sit for WAEC SSCE at government expense. This was to reduce students' failure rate in WAEC SSCE and also to motivate and encourage them to be more serious and diligent in their studies. The performance of students in the State qualifying examination since inception has been encouraging such that, the State has not reneged to set aside sufficient fund annually for adequate preparation, administration, and grading of the answer scripts of the students.

However, despite the huge amount being expended by the state government students' performance in public examinations has been generally unsatisfactory, especially in core subjects such as Mathematics and English Language. Given that the teachers and students have put in efforts in academic preparation because of the high stake attached to the examination, it is important to address the quality of the test items used for the state exam. Therefore, it is pertinent to this study to direct attention towards examining the characteristics of the test items used by the Osun State Ministry of Education to prepare the students for public examinations, more so that there was no evidence that the tests items pass through any standardized testing procedures such validity, reliability, and Differential Item Functioning analysis (which is germane of this study).

Mathematics and English language are the major and pre-requisite subjects for gaining admission into higher institution of learning these days, it is important to examine DIF techniques that can be used to determine the degree to which the two subjects are free of DIF across different groups of examinees. This may be necessary at this time especially considering the major challenges faces by students in passing these subjects. There are many methods for DIF detection proposed over the past two decades. This study focused on Chi-square method because of its strength and power in detecting DIF items. Also, the study adopted Cronbach alpha coefficient and Factor Analysis to ascertain the unidimensionality of Mathematics and English Language of Multiple-choice items of OSQ examination.

2. Statement of the Problem

Differential Item Functioning can lead to an unfair advantage or disadvantage for certain subgroups in educational and psychological testing. There are many competing approaches for the conduct of DIF analyses and many criteria for determining what constitutes significant DIF in items that are scored dichotomously. Although many DIF methods abound, a relatively small number of these methods are preferred based on their theoretical and empirical strengths (Clauser&Mazor, 1998). Three of the preferred methods frequently used to detect item with DIF are Chi-square test, Transformed Item Difficulty, b-Parameter. Literature had shown that there were divergent results on the effectiveness of these three methods based on their strength and power in identifying DIF items.

3. Purpose of the Study

The study was designed to assess the unidimensionality and occurrence of Differential Item Functioning (DIF) inEnglish Language and Mathematics items of the Osun State Qualifying Examination (OSQE). These were with a view to improving the quality of test items to ensure valid decisions.The objectives of the study are to:

a) determine the dimensionality of the items in selected subjects (English Language and Mathematics) of Osun State Qualifying Examination for Senior Secondary School students;

b) establish the occurrence of DIF in the selected subjects of the OSQE;

4. Research Questions

The following research questions were raised from the above stated objective.

1. What is the dimensionality of the OSQ Mathematics Examination?
2. What is the dimensionality of the OSQ English Language Examination?
3. Does DIF exist in Mathematics and English Language items of OSQ examination?

5. Methodology

The research design used was ex-post-facto. The population for the study consisted of all the responses of students who sat for Mathematics and English Language of the Osun State Qualifying Examination in 2008.A sample of 4156 Senior School II students' responses to 50 multiple-choice Mathematics and 50 multiple-choice English Language items of the OSQE for 2008 were used in the study. The sample size for the study selected using purposive and stratified sampling techniques. Research instruments used for the study were adopted 50 multiple-choice Mathematics and 50 multiple-choice English Language items of the OSQE for 2008. The responses of students to 50 multiple-choice Mathematics and 50 multiple-choice English Language items of the OSQE were collected using the Optical Mark Recorder (OMR) sheets. The collected response was coded "1", while the wrong option was coded "0". The reliability coefficients of the 50 multiple-choice English Language and 50 multiple-choice Mathematics questions using Cronbach's Alpha coefficient were found to be 0.87 and 0.88 respectively. Chi-square and Factor Analysis were used to analyse the data using Microsoft excel and SPSS 17.

6. Results

6.1 Research Question One: *What is the Dimensionality of the OSQ Mathematics Examination?*

To answer this question, the responses of the students on 50 multiple-choice items of OSQ Mathematics examination were subjected to factor analysis, as this is a very important step prior to performing DIF analysis. The Cronbach alpha coefficient was found to be 0.85 which showed high internal consistency indicating that the OSQ Mathematics Examination was unidimensional. In the factor analysis, the initial communalities showed the variance in each variable which are accounted for by all components. For principal components extraction, this was equal to 1.0 as the standard rule for correlation analyses. The Extraction communalities showed the estimates of the variance in each variable accounted for by the components. The principal component analysis revealed that the correlation matrix had its entire coefficients less than 0.3. That shows that the item loadings are considered relevant and contributed to the factor loadings as shown is Table one.The extraction from the principal component analysis after interacting of communalities showed thirteen components with eigenvalues greater than 1 as revealed in the Scree plot (see Figure two). This explained 14.066, 5.543, 4.782, 4.250, 3.569, 3.164, 2.986, 2.582, 2.517, 2.402, 2.302, 2.104, and 2.0855% of variance accounted for by each component to the total variance in all of the items. Furthermore, for the 50 multiple-choice Mathematics items, with respect to the eigenvalue greater than 1, the total percentage variance was 52.353.

Table 1. Factor Correlation Matrix of OSQ Mathematics Examination of 50 Multiple-Choice Items

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
F1	1.000												
F2	0.102	1.000											
F3	0.127	0.056	1.000										
F4	0.218	0.023	0.123	1.000									
F5	0.237	0.038	0.122	0.241	1.000								
F6	0.228	0.086	0.098	0.232	0.224	1.000							
F7	0.229	0.085	0.045	0.188	0.144	0.108	1.000						
F8	0.236	0.027	0.120	0.158	0.051	0.272	0.236	1.000					
F9	0.210	0.051	0.155	0.161	0.168	0.222	0.226	0.137	1.000				
F10	0.062	-0.042	-0.029	0.005	0.114	0.029	-0.015	-0.052	-0.002	1.000			
F11	0.231	0.054	0.106	0.037	0.117	0.186	0.111	0.088	0.122	0.080	1.000		
F12	0.169	-0.018	0.054	-0.067	0.007	0.138	0.175	0.184	0.171	0.138	0.158	1.000	
F13	0.107	-0.094	0.134	0.029	0.044	0.112	0.069	0.123	0.041	0.150	0.098	0.229	1.000

From Table 1, it could be seen that the correlation ranges from -0.094 to 0.231which is less than correlation value of 0.3. This showed low correlation value and evidence that OSQ Mathematics Examination is unidimensional. Figure one further confirmed the unidimensionality nature of the examination.

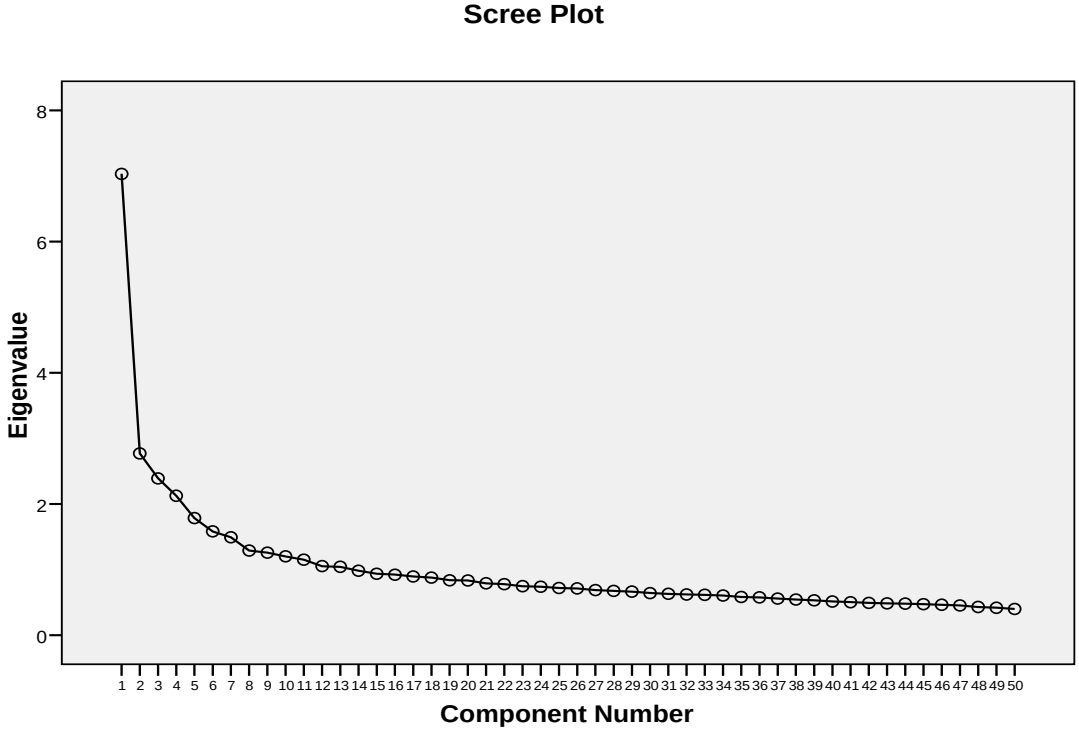

Figure 1. Scree Plot Showing Unidimensionality of Mathematics Items

The Figure 1 is the scree plot for the 50 multiple-choice OSQ Mathematics Examination items. The factor analysis that was performed on the items using extraction method of principal component analysis showed that the first factor having the initial eigenvalue (7.033) which clearly exceeded that of the second factor (2.771) as also revealed in Figure two. From Figure two, the Scree plot showed a visual of the total variance associated with each factor. The steep slope showed the large factors associated with the loading greater than the eigenvalue of 1. The gradual trailing off (scree) showed the rest of the factors lower than an eigenvalue of 1. There are thirteen factors whose values are greater than eigenvalue of 1 and one extracted communality factor distinctly highly than others, showing that the test is unidimensional in nature. Also, it can therefore be concluded that the 50 multiple-choice mathematics items is unidimensional.

6.2 Research Question Two: What is the Dimensionality of the OSQ English Language Examination?

To answer this question, the same procedure used in answering research question one was used. The cronbach alpha coefficient was found to be 0.86 which showed high internal consistency indicating that the OSQ English Language Examination was unidimensional. In the factor analysis, the extraction from the principal component analysis after interacting of communalities showed thirteen components with eigenvalues greater than 1 as revealed in the Scree plot shown in Figure 2.This explained 14.295, 5.537, 4.075, 3.502, 3.195, 3.020, 2.975, 2.698, 2.555, 2.494, 2.404, 2.246, and 2.160% of variance accounted for by each component to the total variance in all of the items. Furthermore, for the 50 multiple-choice Mathematics items, with respect to the eigenvalue greater than 1, the total percentage variance was 51.154.

Table 2. Factor Correlation Matrix of OSQ English Language Examination of 50 Multiple-Choice Items

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
F1	1.000												
F2	-0.225	1.000											
F3	-0.042	0.037	1.000										
F4	-0.053	0.163	0.233	1.000									
F5	0.224	-0.081	0.064	0.052	1.000								
F6	0.222	-0.228	-0.037	-0.031	0.158	1.000							
F7	0.109	0.013	0.031	0.089	0.073	0.201	1.000						
F8	0.085	-0.008	0.113	0.074	0.043	0.180	0.213	1.000					
F9	0.142	0.017	0.043	0.005	0.087	0.091	0.188	0.123	1.000				
F10	0.012	-0.013	0.053	-0.039	-0.070	0.162	0.225	0.203	0.126	1.000			
F11	0.103	-0.009	-0.095	-0.054	-0.047	0.065	0.019	0.036	-0.028	0.053	1.000		
F12	0.092	-0.002	0.063	0.009	-0.026	0.224	0.218	0.211	0.115	0.014	0.119	1.000	
F13	0.055	0.008	0.070	0.039	0.036	0.142	0.168	0.210	0.087	0.052	0.047	0.199	1.000

From Table 2, it could be seen that the correlation ranges from -0.095 to 0.228which is less than correlation value of 0.3. This showed low correlation value and evidence that OSQ English Language Examination is also unidimensional. Figure three confirmed the unidimensionality nature of the examination.

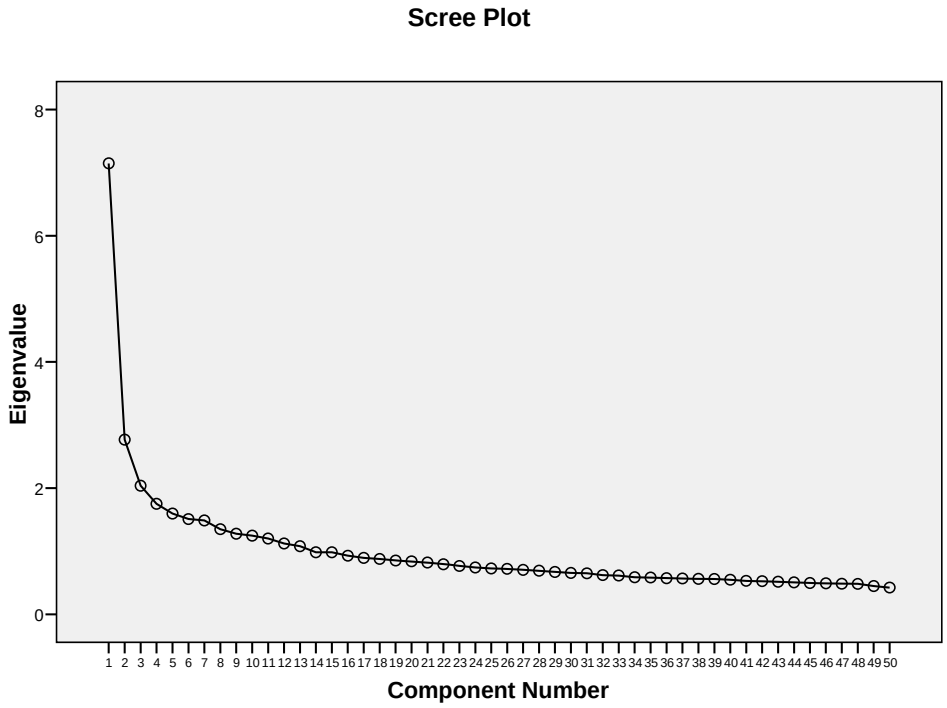

Figure 2. Scree Plot Showing Unidimensionality of English Language Items

Figure 2 is the scree plot for the 50 multiple-choice OSQ English Language Examination items. From Figure three, the Scree plot showed a visual of the total variance associated with each factor. The steep slope showed the large factors

associated with the loading greater than the eigenvalue of 1. The gradual trailing off (scree) showed the rest of the factors lower than an eigenvalue of 1. There are thirteen factors whose values are greater than eigenvalue of l and one extracted communality factor distinctly highly than others, showing that the test is unidimensional in nature. The factor analysis that was performed on the items using extraction method of principal component analysis (see appendix 4) showed that the first factor having the initial eigenvalue (7.147) clearly exceeded that of the second factor (2.768) as also revealed in Figure 3. This also concluded that the 50 multiple-choice English language items is unidimensional.

6.3 Research Question Three: Does DIF Exist in Mathematics and English Language Items of OSQ Examination?

To answer this question, chi-square method with 0.05 level of significant was used to establish the presence of DIF in both Mathematics and English Language items.

Table 3. Summary of Results from the Chi-square Method of Detecting Differential Item Functioning OSQ Mathematics and English Language Examinations

Item	Mathematics χ^2	Significant level	Enlish Language χ^2	Significant level
1	0.43	0.51	0.74	0.39
2	0.13	0.77	9.32*	0.00
3	7.39*	0.01	1.46	0.23
4	8.83*	0.00	0.37	0.55
5	22.55*	0.00	1.06	0.30
6	0.69	0.41	1.36	0.24
7	0.91	0.34	0.02	0.90
8	3.26	0.07	1.79	0.18
9	0.32	0.57	0.87	0.35
10	0.65	0.42	0.54	0.46
11	0.22	0.64	3.38	0.07
12	0.24	0.63	1.03	0.31
13	7.80*	0.01	0.00	0.98
14	1.99	0.16	4.38*	0.04
15	1.81	0.18	4.14*	0.04
16	0.29	0.59	0.02	0.88
17	7.34*	0.01	5.79*	0.02
18	3.71	0.06	0.12	0.72
19	0.25	0.88	0.53	0.47
20	3.66	0.06	1.47	0.23
21	3.24	0.07	2.39	0.12
22	2.57	0.11	2.91	0.09
23	12.55*	0.00	0.14	0.70
24	2.92	0.09	7.62*	0.01
25	7.01*	0.01	0.06	0.80
26	0.06	0.81	1.37	0.24
27	15.06*	0.00	0.37	0.54
28	7.47*	0.01	4.10*	0.04
29	0.01	0.91	0.55	0.46
30	0.04	0.84	4.54*	0.03
31	1.54	0.21	0.45	0.50
32	0.48	0.49	3.18	0.08
33	0.95	0.33	0.03	0.86
34	7.93*	0.01	0.13	0.71
35	0.10	0.75	0.34	0.56
36	6.62*	0.01	1.41	0.24
37	2.55	0.11	1.74	0.19
38	0.75	0.39	0.68	0.41
39	0.63	0.43	4.92*	0.03
40	0.48	0.49	0.52	0.47
41	15.64*	0.00	0.02	0.90

42	2.67	0.10	30.00*	0.00
43	0.91	0.34	0.17	0.68
44	0.01	0.92	0.12	0.73
45	0.88	0.35	3.41	0.07
46	1.20	0.27	4.99*	0.03
47	3.89	0.06	0.18	0.67
48	0.10	0.76	2.75	0.10
49	16.21*	0.00	0.58	0.45
50	10.98*	0.00	1.09	0.30

*Item reveals DIF ($p \leq 0.05$)

Table 3 showed items that flag DIF. For item to flag DIF, the chi-square significant value must be less than 0.05. When this is applied to Mathematics items, the chi-square significant value less than 0.05 procedure flagged fourteen items representing 28% of the 50 items as displaying DIF (the items: 3, 4, 5, 13, 17, 23, 25, 27, 28, 34, 36, 41, 49, and 50). Also, for English Language items, the chi-square procedure flagged 10 items representing 20% of the 50 items as displaying DIF (2, 14, 15, 17, 24, 28, 30, 39, 42, and 46). It can therefore be concluded that both Mathematics and English Language items exhibited DIF items.

7. Discussion

Differential Item Functioning analysis is recommended only when the test scores are unidimensional (Clauser,&Mazor 1998). There are various ways for testing unidemensionality (Tate, 2003). However, unidimensionality can be established when one of two conditions is met from the results of an exploratory factor analysis (Reckase, 1999): first, a factor analysis on the inter-item correlation matrix should show that the first factor accounts for at least 20% of the variance of the unrotated factor matrix or second the eigen value of the first factor should clearly exceed that of the second factor. In another study by Wiberg (2004) a high cronbach alpha coefficient approach was used to indicate unidimensionlity. In the same vein, Norusis (2004) postulated statistical independence among variables to confirm its unidimensionality. The results of the research questions 1 and 2 as revealed in Tables 7 and 8, and subsequently Figures 2 and 3 showed that both subjects (Mathematics and English Language) are evidences of unidimensionality.The third research question was based on establishing the occurrence of DIF in the OSQ Mathematics and English Language examination. Mantel Haezel Chi-square method was used in detecting the occurrence of DIF in the two subjects being one of the most popularly used methods (Nabeel 2010). The results of this study, alone and in combination with Maller's (2001) showed that significant numbers of items on OSQ Mathematics and English Language examination displayed DIF.

8. Conclusion

The study concluded that each of the multiple-choice Mathematics and English Language items administered by the Osun State Ministry Education measured a single construct which showed evidence of unidimensionality.The study also revealed that both Mathematics and English Language items exhibited DIF items.The study implied that undimensionality of test items is a necessary condition for DIF analysis. It also implied that the detection DIF in multiple-choice itemswill help test developers to generate quality items that will subsequently ensure correct interpretations of test scores.Test practitioners should endeavour to perform DIF analysis from a pilot study before administration of test(s) so that items that function differently for different test taking groups can be identified for possible replacement.Future studies may consider science oriented subjects for possible manifestation of DIF.

9. Acknowledgements

I sincerely appreciate efforts and assistance I got from the officials of the Osun State Ministry of Education for making available the data used for this study.

References

Clauser, B.E., & Mazor, K.M. (1998). Using Statistical Procedure to Identify Differentially Functioning Test Items. *Educational Measurement: Issue and Practice, 17,* 31-44. http://dx.doi.org/10.1111/j.1745-3992.1998.tb00619.x

Green, K. E. (1996). Applications of the Rasch model to evaluation of survey data quality. *New Directions Evaluation, 70,* 81-92. http://dx.doi.org/10.1002/ev.1036

Lee, J. (1997). State activism in education reform: Applying the Rasch model to measure trends and examine policy coherence. *Educational Evaluation and Policy Analysis, 19*(1), 29-43. http://dx.doi.org/10.3102/01623737019001029

Nabeel, A. (2010). A Gender-Related Differential Item Functioning of Mathematics Test Items. *The International Journal of Educational and Psychological Assessment, 5,* 101-116.

Norusis, M. (2004). *SPSS 13.0 Statistical Procedures Companion*. Upper Saddle-River, N.J.: Prentice Hall, Inc..

Reckase, M. D. (1999). Uni-factor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 14*(2), 23 – 29.

Tate, R. (2003). A Comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*, 159-203. http://dx.doi.org/10.1177/0146621603027003001

Wiberg, M. (2004). Classical Test Theory versus Item Response Theory. An Evaluation of the Theory of Test in the Sweden Driving-License Test. Retrieved from www.umcl.se