ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 3, No. 3, pp. 401-412, May 2012
© 2012 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/jltr.3.3.401-412

Issues Affecting Item Response Theory Fit in Language Assessment: A Study of Differential Item Functioning in the Iranian National University Entrance Exam

Alireza Ahmadi
Shiraz University, Iran
Email: arahmadi@shirazu.ac.ir

Nathan A. Thompson
Assessment Systems Corporation, USA

***Abstract*—This study aimed at examining the issues affecting the use of IRT models in investigating differential item functioning in high stakes testing. It specifically focused on the Iranian National University Entrance Exam (INUEE) Special English Subtest. A sample of 200,000 participants was randomly selected from the candidates taking part in the INUEE 2003 and 2004 respectively. The data collected in six domains of vocabulary, grammar, word order, language function, cloze test and reading comprehension were analyzed to evaluate the applicability of item response theory (IRT; Embretson & Reise, 2000), including the use of IRT for assessing differential item functioning (DIF; Zumbo, 2007). Substantial model-data misfit was observed in calibrations using PARSCALE and BILOG MG software (Scientific Software International, 2004). Additional analysis through Xcalibre and Iteman 4 (Assessment Systems Corporation, 2010) suggested that item response theory, including IRT-based DIF analysis, is not applicable when the test administered is noticeably beyond the participants' level of capability, when the test is speeded, or if students are penalized for their wrong answers.**

***Index Terms*—IRT, DIF, Iranian National University Entrance Exam**

I. Introduction

Item Response Theory (IRT), also called latent trait theory, is the most popular modern test theory which has attracted lots of attention and is considered as an active area of research in the world of assessment and testing. Item response theory is a mathematical model that specifies the relation of trait levels and item characteristics to a person's item response (Embretson & Riese, 2000). Hambleton et al, (1991) state that:

IRT rests on two basic postulates: a) the performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits or abilities; and b) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (p.7).

IRT is more complex than its classical counterpart, classical test theory (CTT), since it requires more assumptions and the use of special software, not many of which are adequately user-friendly for the majority of those interested in assessment and testing. However, it can explain a lot of things for which the classical test theory has either no explanation or provides weaker and less accurate justifications. CTT is based on the assumption that a test-taker's observed score is a combination of his true score and the error score. It requires weaker assumptions and therefore is relatively easy to interpret. Because of that it is still very common in the world of testing.

However, IRT offers many important advantages over CTT. Henning (1987) mentions the advantages as: sample-free item calibration, test-free person measurement, multiple reliability estimation, identification of guessers and other deviant responders, potential ease of administration and scoring, economy of items, test tailoring facility, test equating facility, item banking facility, reconciliation of norm-referenced and criterion-referenced testing, item and person fit validity measures, score reporting facility, the study of item and test bias, and the elimination of boundary effects in program evaluation. Although some of these features are also present in CTT, IRT provides a better index of each of these. Through IRT one can also compare different test takers who have taken different versions of a test (Hambleton & Swaminathan, 1985).

IRT is based on a number of assumptions. First of all, it assumes uni-dimensionality; that is, the test measures only one latent trait which is usually referred to as 'ability,' denoted by θ. An entwined assumption is the concept of local independence; that is the item responses are assumed to be independent of one another. The assumptions of unidimensionality and local independence are related in that; “items found to be locally dependent will appear as a separate dimension in a factor analysis” (Reeve 2003, p. 12). Besides factor analysis, model fit can provide evidence that this assumption is satisfied.

The second assumption is that the probability of a certain response to an item is a function of θ, and can be mathematically modeled. There are numerous mathematical models available, both for dichotomous (correct/incorrect) data and polytomous (rating scale or partial credit) data. An evaluation of model-data fit is essential for providing evidence that this assumption is satisfied.

The invariance assumption is the third assumption, which states that the item parameters are not influenced by the sample characteristics. This means that unlike classical test theory where parameter estimates and statistics vary across samples, item parameters are considered invariant to group membership in IRT. This is a great advantage of IRT. “The property of invariance of ability and item parameters is the cornerstone of IRT. It is the major distinction between IRT and classical test theory” (Hambleton, 1994, p. 540).

There are three IRT models in widespread use for dichotomous data, all of which require the above-mentioned assumptions. The simplest IRT model is a one-parameter logistic (1PL) model, a version of which is also known as the Rasch model. It is based on the item parameter *b* (item difficulty). The difficulty is the value of ability when a person has a 50% probability of answering an item correctly. Usually, the difficulty is standardized and ranges from -3 to +3 with higher values indicating more difficult items. This model assumes that all the items are equally discriminating. The two-parameter model is an extension of the 1PL as it adds an item discrimination parameter (*a*) to the model. The discrimination parameter determines how well an item discriminates between persons with high and low ability. This parameter affects the steepness of the item characteristic curve (ICC); as its value increases the slope of ICC increases. Usually, the discrimination parameter ranges from 0 to 2. The three-parameter logistic (3PL) model extends the 2PL model by including a pseudo-guessing parameter. This parameter estimates the probability of answering an item correctly for persons having very low ability. Adding this parameter to the model results in the lower asymptote of the ICC being nonzero, typically $1/k$ where k is the number of item options. This differs from the 1PL and 2PL models where persons of very low ability have a zero probability of answering the item correctly.

The potential of IRT for solving different kinds of testing problems is considerable, provided that there is fit between the model and the test data. IRT is applied to the investigation of many issues, including DIF & item bias analysis, test linking and equating, adaptive testing, program evaluation and assessment testing, and test assembly.

The present study is related to the use of IRT in studying DIF in high-stakes tests. IRT can provide a theoretically useful tool for DIF analysis since DIF can be modeled through the use of estimated item parameters and ability. In fact, DIF is very often studied in the context of item response theory. DIF occurs when the responses provided by students of approximately equal ability are significantly different based on students’ membership in a particular subgroup. In other words, respondents with similar ability levels from different populations, have a different probability of responding to an item (Camilli & Shephard, 1994).

Differential item functioning methods allow one to judge whether items (and ultimately the test they constitute) are functioning in the same manner in various groups of examinees. In broad terms, this is a matter of measurement invariance; that is, is the test performing in the same manner for each group of examinees? (Zumbo, 2007, p. 1).

DIF items are usually considered as serious threats to the validity of the instruments measuring the ability levels of individuals from different groups. Such instruments cannot be considered as sufficiently valid for between-group comparisons, as their scores may be indicative of a variety of attributes other than the ones the scale is intended to measure (Thissen, Steinberg, & Wainer, 1988). Thus DIF detection is a crucial step for all testing situations. It becomes “intimately tied” to test validation to establish the inferential limits of the test; that is, whether the inferences made on the test scores are valid for a specific group (Zumbo, and Rupp, 2004; and Zumbo, 2007). In case of high-stakes tests, DIF analysis is of higher importance and becomes compulsory (Pae & Park 2006).

IRT methods of studying DIF are based on comparing the ICCs between groups (Embretson & Reise, 2000). This is the same as comparing the item parameter estimates for persons of matched ability. There are several IRT approaches for DIF detection. For example, some use the area between the ICCs (e.g., Raju, 1988); some use statistical testing of the equality of the ICC parameters (e.g., Lord, 1980); and others use statistical testing of the model fit (e.g., Thissen, Steinberg, & Wainer, 1988).

Analyses based on the one-parameter logistic IRT model, or the Rasch dichotomous model, investigate DIF in the threshold or location parameter *b*. They test whether the reference and focal groups have a different probability of answering an item correctly after controlling for group differences on the latent variable? This method has strict requirements for the Rasch model to keep its elegance (e.g., sum score sufficiency). Any item that differs from the other items in its ability to discriminate among respondents is considered a misfitting item to the Rasch model (Smith, 1991). Thus, if an item has different estimated slopes (i.e., discrimination ability) between the reference and focal groups, the item is considered misfit and is usually eliminated.

Some researchers (e.g., Angoff, 1993; Camilli & Shepard, 1994) believe that investigation of DIF in the framework of Rasch measurement is limited. Exclusion of the discrimination power or pseudo-guessing as the possible sources of DIF will result in undetected DIF items and may hence lead to the removal of the most useful items in a measure (Angoff, 1993). Therefore, applications of the Rasch models limit our understanding of the group differences in responding to items. As such, IRT models that allow the discrimination parameter to vary from item to item describe the data more accurately than the ones that limit the slope parameter to be equal across items.

For binary data, the two-parameter logistic IRT model studies DIF in relation to the item's threshold parameter b, slope parameter a, or both parameters. DIF in the slope parameter represents an interaction between the underlying measured variable and group membership (Teresi, Kleinman, & Ocepek-Welikson, 2000). The degree to which an item represents the underlying construct depends on the group being measured.

The 3-parameter model allows for the investigation of DIF in the discrimination parameter, threshold parameter, and the pseudo-guessing parameter.

II. Iranian National University Entrance Examination

The Iranian National University Entrance Exam (INUEE) is designed to screen candidates for studying at higher education. It is given to high school graduates who intend to continue their studies at the university level. The INUEE consists of two parts. The first part, the general part, is designed to measure applicants' general academic ability, and focuses on subjects such as Islamic studies and culture (theology), Persian language and literature, Arabic language, and one elective foreign language (English, French, German, Italian, or Russian). It is believed that these subjects play a disproportionate role in applicants' overall academic ability; hence the scoring system which is used is a weighted one in the sense that e.g. a correct response to an item of Islamic studies and culture is considered more important than a correct response to an item of the Arabic language. The general part of the INUEE includes 100 MC items with 25 items dedicated to each subject area. This part of the test is the same in the subjects, number and form of the items for all the applicants independent of their high school majors. However, the content of the items usually differs.

The second part of the test, the special part, focuses on subjects related to the four high school majors of the applicants in mathematics, natural sciences, humanities, and arts. Students are admitted to different fields of study in higher education depending on their score in the first and second part of the test altogether. This part includes 70-150 MC items depending on students' major in high school. The subject areas and the content of the items are also determined according to the majors. Like the first part, a weighted system is used to score the items in each subject area. The INUEE is a competition test and the best candidates are selected for the limited number of vacancies available for each field of study in different universities.

The applicants are ranked on the basis of their total scores on both parts and admitted to the universities in the majors they had requested. If an applicant's score is not high enough to be admitted to their requested discipline, the applicant can be admitted to other disciplines. Although many applicants are not accepted in their majors of interest, they may still continue because in addition to the social desirability of getting into universities, male students are exempt from compulsory military service (Farhady & Hedayati, 2009, p.136).

The second part of the test is administered in 5 subtests over three days, with each subtest being administered in half a day. Four subtests are related to the four high school majors in Iran (mathematics, natural sciences, humanities, and arts) and the fifth subtest is specially designed for those applicants whose intended university major is English or other foreign languages. Each high school graduate can sit for up to 3 subtests to earn acceptance in different fields of studies in universities. Applicants can take only one of the subtests related to the mathematics, natural sciences and humanities major. They are also allowed to sit for the other two versions related to the arts and foreign languages if they like.

III. Purpose of the Study

The present study aimed at finding the issues affecting the use of IRT models in investigating differential item function in high stakes test. It specifically focused on the INUEE Special English Subtest.

IV. Methodology

A. Participants

The data for this study came from 200,000 participants randomly selected from among more than 500,000 high school graduates who sat for the Special English Subtest of the Iranian National University Entrance Exam in 2003 and 2004 respectively. There were 270,201 examinees in 2003 and 284,079 examinees in 2004; 100,000 were selected from each sample.

B. Instrument

Iranian National University Entrance Exam (Special English Subtest)

The foreign language subtest of this test that taps candidates' knowledge of grammar and lexicon as well as general reading comprehension has two parts. The Special English Subtest, plays a more important role in applicants' admission to universities in foreign language studies and that is why it was selected for investigation in the present study. This test consists of 70 MC items in six areas of language: structure (10-12 items), vocabulary (20 items), word order (4-5 items), language function (4-5 items), cloze test (15 items), and reading comprehension (15 items).

C. Data Collection Procedure

The data of this study was collected through the kind cooperation of the Center of Educational Measurement. It provided the researchers with the answer sheets of all the participants taking the INUEE Special English Subtests 2003 and 2004.

D. Data Analysis

The data of the study were subjected to CTT analysis using *Iteman 4* and IRT analysis using PARSCALE and *Xcalibre*, including DIF detection using PARSCALE and BILOG. Because the test consists of multiple choice items, the three-parameter model (Embretson & Reise, 2000) was utilized.

V. RESULTS AND DISCUSSION

The classical analysis of the items using *Iteman 4* indicated that the test had adequate reliability but was quite difficult. Tables 1 presents the summary statistics of the 2003 test, for all the items, and for each domain (content area).

TABLE 1:
SUMMARY STATISTICS FOR DIFFERENT DOMAINS OF THE INUEE ENGLISH SUBTEST (2003)

Score	Items	Mean	SD	Var.	Min Score	Max Score	Alpha	Mean P	Mean r_{pbis}
All items	70	25.636	12.389	153.488	0.00	69.00	0.912	0.481	0.293
Domain 1 (Voc.)	10	3.676	1.957	3.829	0.00	10.00	0.455	0.432	0.221
Domain 2 (Gram,)	20	8.650	4.635	21.486	0.00	20.00	0.834	0.536	0.349
Domain 3 (W.O.)	5	2.431	1.487	2.210	0.00	5.00	0.648	0.584	0.324
Domain 4 (Lg Func)	5	1.762	1.362	1.856	0.00	5.00	0.495	0.460	0.313
Domain 5 (Cloze)	15	5.084	2.962	8.775	0.00	15.00	0.678	0.466	0.239
Domain 6 (R.C.)	15	4.032	3.027	9.164	0.00	15.00	0.680	0.430	0.304

As indicated in this table, the reliability coefficient was 0.912, but the mean score was 25.636 out of 70 (36.61%) and the mean *P* was 0.481, which is very low performance for a national test. Note that the mean score is a better representation of test difficulty because the mean *P* does not include omitted responses. This low performance is observed in all the six domains. The best performance is seen in domain 3 (48.62%) and the lowest performance is seen in domain 6 (26.88%). The level of performance for the other four domains are as follow: domain 1: 36.76%, domain 2: 34.25%, domain 4: 35.24%, and domain 5: 33.89%. Figure 1 depicts the same results more clearly by displaying the distribution of the total number correct scores.

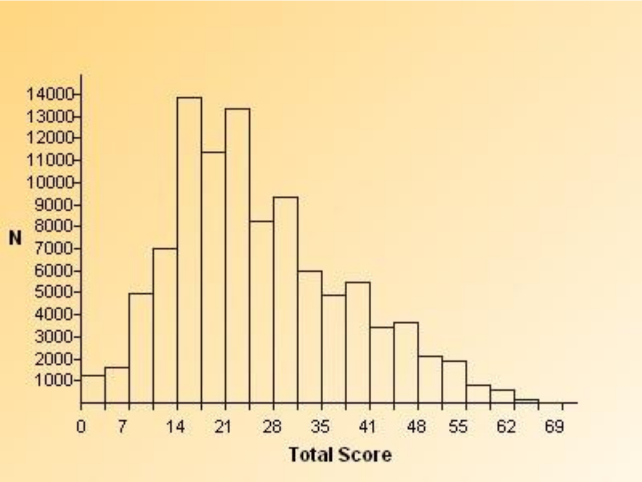

Figure 1: Total scores for different domains of the INUEE English Subtest (2003)

Table 2 presents the summary classical statistics of the 2004 test, and confirms the results of Table 3. Here again it is found that while the reliability is 0.901, the overall performance is very low with the mean of 21.63 (30.91%), which is even lower than the performance on the test 2003. The best performance is seen in domain 4 (language function with 35.84%, accuracy) and the lowest performance is seen in domain 6 (reading comprehension with 28.71%.accuracy). The accuracy level of performance for the other four domains is as follows: domain 1 with the accuracy of 33.77%, domain 2 with the accuracy of 30.15%, domain 3 with the accuracy of 32.5%, and domain 5 with the accuracy of 30.03%.

TABLE 2:
SUMMARY STATISTICS FOR DIFFERENT DOMAINS OF THE INUEE ENGLISH SUBTEST (2004)

Score	Items	Mean	SD	Var.	Min Score	Max Score	Alpha	Mean P	Mean r_{pbis}
All items	70	21.636	11.757	138.233	0.00	67.00	0.901	0.445	0.271
Domain 1 (Voc.)	10	3.377	2.190	4.796	0.00	10.00	0.591	0.426	0.273
Domain 2 (Gram,)	20	6.030	3.738	13.974	0.00	20.00	0.720	0.416	0.256
Domain 3 (W.O.)	5	1.625	1.206	1.453	0.00	5.00	0.370	0.446	0.331
Domain 4 (Lg Func)	5	1.792	1.270	1.612	0.00	5.00	0.420	0.516	0.216
Domain 5 (Cloze)	15	4.504	2.964	8.783	0.00	15.00	0.685	0.437	0.246
Domain 6 (R.C.)	15	4.307	3.435	11.796	0.00	15.00	0.755	0.481	0.313

Figure 2 displays the results presented in the first row of Table 2 more clearly. It displays the distribution of the total number correct scores.

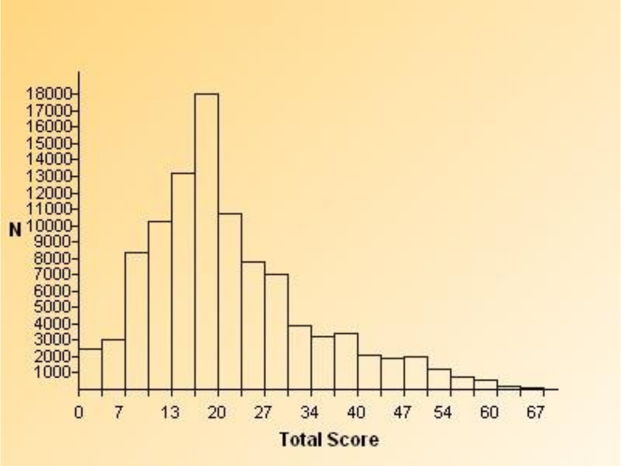

Figure 2: Total scores for different domains of the INUEE English Subtest (2004)

Both tests were calibrated with the three-parameter logistic IRT model (3PL). With the 3PL, the probability of an examinee with a given θ correctly responding to an item is (Hambleton & Swaminathan, 1985, Eq. 3.3):

$$P_i(X_i = 1 | \theta_j) = c_i + (1-c_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1+\exp[Da_i(\theta_j - b_i)]} \qquad (7)$$

where
a_i is the item discrimination parameter,
b_i is the item difficulty or location parameter,
c_i is the lower asymptote, or pseudoguessing parameter, and
D is a scaling constant equal to 1.702 or 1.0.

The *a* parameter ranges in practice from 0.0 to 2.0, with a higher value indicating more discriminating power. The *b* parameter typically ranges from -3 to +3, as it is indicative of the examinee ability level for which the item is appropriate on a scale that is analogous to the standard normal scale. The *c* parameter is typically near 1/*k*, where *k* is the number of alternatives to a multiple choice item. The INUEE test is composed of four-option items, so this value can be expected to be approximately 0.25 on average.

IRT calibrations were completed with both PARSCALE and *Xcalibre*. Detailed results are presented in Appendices A and B, while summary results are presented in Table 3. As with classical analysis, items had strong discriminations but were extremely difficult. The mean *b* parameters were 1.38 and 1.03 with PARSCALE, and 1.31 and 1.55 with *Xcalibre,* all of which imply that the average item is appropriate for a student in the top 15% of the population. This result is even more notable when considering that more than 25% of the responses were omitted in 2003 and more than 32% in 2004; had examinees been required to answer each question, items would appear even more difficult.

TABLE 3:
SUMMARY STATISTICS OF IRT CALIBRATION

		PARSCALE			Xcalibre					
Statistic	*a*	*b*	*c*	*a*	*b*	*c*	*R*	*P*	r_{pbis}	Omit
2003 mean	0.90	1.38	0.21	1.11	1.31	0.23	4.53	0.45	0.27	25650
2003 SD	0.50	3.80	0.16	0.55	1.06	0.10	7.51	0.16	0.14	11870
2004 mean	0.97	1.03	0.20	1.27	1.55	0.25	3.29	0.45	0.27	32198
2004 SD	0.40	1.61	0.08	0.49	0.95	0.08	4.62	0.16	0.14	12652

While IRT provides many advantages in test development and analysis, an essential criterion to its application is acceptable model fit. Finally, a likelihood-ratio -statistic for each item is computed by

$$\chi 2j = 2\sum_{h=1}^{H} r_{hj} \ln \frac{r_{hj}}{N_{hj} P_j(\theta_h)}$$

where H_j is the number of intervals for item *j* and r_{hj} is the observed frequency for interval *h* in item *j*. The degree of freedom is the number of response categories minus 1; since all items on the INUEE are dichotomously scored, this is always 1. *Xcalibre,* in contrast, standardizes the residual. Therefore, a value greater than 1.96 indicates a rejection of fit with a significance of 0.05.

Item fit statistics were quite poor. Every single item on both tests was rejected with PARSCALE's chi-square fit statistics. With *Xcalibre's* standardized residual fit statistics, 61 items were rejected for the 2003 test and 45 items for the 2004 test. The average residuals are reported in Table 3; average chi-square statistics from PARSCALE could not be calculated because many were too large to be included in output. Some of the worst fitting items were eliminated in an iterative attempt to improve the data-model fit, but most items continued to be rejected.

Such extensive misfit is likely caused by additional variables affecting the process of responding to items; IRT assumes that the probability of correctly responding is a function only of θ. Three factors were speculated for such a misfit two of which are related to the substantial number of omitted responses seen in Table 3. First, the test could have been too speeded; examinees did not have sufficient time to respond to items according to their ability. Secondly, students were penalized for their wrong answers; every three wrong answers will cancel a correct answer on this test. This correction for guessing on the INUEE discourages many students from responding to all items, and as such their performance is underestimated. Finally, such misfit could be due to the fact that the items were too difficult for the target population, leading to the skewed raw score distributions in Figures 1 and 2.

The model misfit substantially inhibited the investigation of DIF using IRT. PARSCALE, like BILOG-MG, characterizes DIF as different item parameters for relevant groups. It then calculates two significance tests for the comparison, the more conservative of which is a chi-square test. For this study, the *a* and *c* parameters were held constant, and the *b* parameter allowed to vary, which evaluates whether there was differential difficulty between the two gender groups. As seen in Appendices C and D, most items were rejected for DIF, and nearly every item that was not was a case where PARSCALE was not able to estimate parameters and a *b* parameter of 0.00 was supplied instead. It is unlikely that nearly every item would be rejected for DIF, suggesting that the fit issues prevented the application of IRT to investigate DIF.

Using other DIF detection softwares did not solve the problem either. BILOG MG was used to see whether the IRT models would fit the data. BILOG MG provides a large-sample test of the goodness-of-fit of individual test items in the analysis. Almost all the items indicated misfit no matter which IRT model was used.

VI. Conclusions

This study was designed to evaluate the presence of DIF on different years of the INUEE English Subtest. It however, became a study of the factors affecting item response theory fit in language assessment after IRT calibrations displayed substantial misfit. The use of IRT models indicated a high level of misfit for almost all the items, precluding effective DIF analysis. This was the case for both PARSCALE and BILOG MG software. Analysis of the results led the researchers to instead evaluate possible causes of this misfit in a 70-item test with a large sample (100,000 students). Plausible causes were speculated to be the difficulty of the test, the speededness of the test, and a scoring penalty for guessing. The existence of the three speculations were confirmed through further analysis. Unfortunately, the existence of all three issues prevents the isolation of any as the cause for misfit. Future research is necessary to investigate this further.

Overall, the study can lead one to conclude that although in many applications IRT is preferable to its counterpart, CTT, it can turn out to be quite inefficient under certain conditions. The present study concluded that it cannot be used for DIF analysis (though it is the most preferred method in the literature) when the test administered is noticeably beyond the participants' level of capability or when the test is speeded, when some students are not able to finish the test on time. A similar problem is present if students are penalized for their wrong answers and this may mean that tests which allow for guessing are preferable to tests in which guessing is suppressed.

Within the context of large-scale language assessment, these results have important implications regarding application of IRT for test development or analysis. It is recommended that the test developers ensure that the effect of speededness is minimized, to ensure that the test is a power test. Additionally, a guessing penalty is likely to inhibit the application of IRT because it violates the unidimensionality assumption of IRT, so it is recommended that such penalties not be applied.

APPENDIX A: 2003 IRT PARAMETERS AND STATISTICS

	Parscale			Xcalibre						
Item	*a*	*b*	*c*	*a*	*b*	*c*	*R*	*P*	r_{pbis}	Omit
1	0.12	-0.54	0.53	1.78	2.74	0.39	2.03	0.40	0.02	21721
2	1.51	0.80	0.28	1.37	0.75	0.28	1.29	0.51	0.41	9258
3	1.51	2.45	0.37	1.67	2.24	0.37	3.09	0.40	0.11	12324
4	0.83	1.84	0.41	1.28	1.81	0.44	4.25	0.50	0.18	13529
5	0.63	0.50	0.26	0.70	0.99	0.31	1.01	0.57	0.30	22251
6	0.80	0.87	0.14	0.76	0.89	0.12	5.88	0.43	0.35	13899
7	0.39	-0.84	0.05	0.48	0.09	0.19	2.60	0.66	0.25	16298
8	0.84	1.86	0.20	1.58	2.52	0.14	8.91	0.16	0.10	18966
9	0.36	3.24	0.12	0.47	3.00	0.15	2.23	0.25	0.14	12505
10	0.94	0.94	0.20	0.94	0.87	0.20	1.52	0.45	0.36	7845
11	1.53	2.17	0.16	1.53	2.16	0.16	4.69	0.21	0.19	23153
12	0.75	0.31	0.28	0.78	0.40	0.29	1.54	0.62	0.33	11292
13	2.41	2.61	0.31	2.35	2.29	0.30	3.72	0.32	0.07	10683
14	0.54	0.08	0.07	0.56	0.68	0.13	3.86	0.54	0.31	21537
15	1.46	0.78	0.16	1.43	0.74	0.16	1.30	0.45	0.48	16083
16	1.31	-0.59	0.20	1.44	-0.54	0.20	2.67	0.77	0.39	8002
17	0.09	0.00	0.00	0.64	-0.68	0.13	4.13	0.78	0.25	14916
18	0.65	-0.14	0.05	0.81	0.38	0.11	4.98	0.60	0.36	26099
19	0.22	-2.19	0.00	0.46	0.84	0.06	6.63	0.49	0.24	22243
20	1.69	1.48	0.12	1.57	1.65	0.12	2.68	0.25	0.40	32162
21	1.17	0.60	0.22	1.14	0.77	0.22	1.04	0.53	0.42	26160
22	1.15	1.25	0.17	1.08	1.22	0.17	1.65	0.36	0.39	14718
23	0.49	0.38	0.12	0.64	1.02	0.18	2.56	0.53	0.30	31302
24	0.91	-0.27	0.09	0.92	-0.04	0.09	5.66	0.65	0.39	17460
25	1.43	0.39	0.20	0.20	3.00	0.41	39.77	0.57	0.49	21013
26	0.57	1.07	0.15	0.66	1.33	0.17	1.13	0.43	0.31	23657
27	1.43	0.54	0.27	0.20	3.00	0.41	35.12	0.57	0.46	19468
28	1.16	-0.52	0.13	1.19	-0.35	0.15	4.16	0.74	0.39	13258
29	0.93	0.01	0.22	1.06	0.51	0.27	1.82	0.66	0.40	31237
30	1.19	0.11	0.27	1.25	0.16	0.26	1.62	0.66	0.41	15892
31	0.73	-0.32	0.09	0.82	-0.13	0.09	5.63	0.65	0.35	13805
32	1.17	-0.13	0.23	1.17	0.01	0.23	2.24	0.69	0.41	17395
33	1.06	2.30	0.18	1.19	2.33	0.18	0.82	0.24	0.18	31856
34	1.01	-0.08	0.15	1.08	0.04	0.15	4.34	0.65	0.41	17035
35	0.09	0.00	0.00	0.63	-0.39	0.06	10.42	0.69	0.27	13188
36	0.37	2.98	0.25	0.48	3.00	0.27	1.79	0.38	0.12	33948
37	1.16	-0.25	0.27	1.33	0.05	0.33	1.87	0.73	0.40	19610
38	0.71	0.55	0.14	0.92	0.86	0.20	0.95	0.51	0.38	24708
39	1.46	1.04	0.17	1.42	0.97	0.16	2.37	0.40	0.44	18825
40	0.62	2.22	0.15	0.72	2.15	0.15	3.36	0.28	0.23	22853
41	1.22	0.26	0.16	0.20	3.00	0.42	39.92	0.57	0.48	17502
42	1.07	2.57	0.17	1.07	2.62	0.18	5.15	0.22	0.14	31451
43	1.06	0.88	0.19	0.92	1.32	0.19	2.74	0.44	0.43	36342
44	0.02	1.69	0.89	2.50	3.00	0.14	9.08	0.15	-0.07	28277
45	0.84	0.76	0.32	1.05	1.19	0.37	3.39	0.56	0.33	31029
46	0.50	0.54	0.10	0.63	1.01	0.18	2.43	0.49	0.32	21132
47	0.10	0.00	0.00	0.35	-0.31	0.16	2.68	0.72	0.14	19269
48	0.18	1.80	0.24	1.22	2.37	0.49	4.19	0.52	0.08	23187
49	1.13	-0.15	0.25	1.14	0.16	0.31	2.60	0.69	0.42	17652
50	0.80	1.86	0.26	0.94	1.93	0.27	2.68	0.38	0.23	28012
51	0.94	0.64	0.31	1.11	0.91	0.34	1.04	0.57	0.35	22689
52	0.59	-0.24	0.04	0.68	0.80	0.21	0.84	0.59	0.36	32345
53	0.91	-0.31	0.14	0.97	0.69	0.31	3.65	0.66	0.44	36923
54	1.00	27.49	0.20	2.50	3.00	0.24	4.29	0.24	0.00	50484
55	0.02	0.00	0.97	2.50	2.86	0.18	8.59	0.18	-0.06	42610
56	1.00	15.95	0.21	2.17	3.00	0.21	2.07	0.21	0.02	28427
57	0.98	1.99	0.18	1.19	2.23	0.19	0.87	0.27	0.26	46282
58	2.21	1.23	0.19	2.20	1.52	0.20	2.43	0.36	0.48	50733
59	0.91	-0.09	0.06	1.06	0.47	0.16	2.06	0.60	0.47	27467
60	0.53	2.48	0.19	0.89	2.43	0.21	0.49	0.31	0.22	51852
61	0.87	-0.58	0.22	1.02	0.22	0.41	3.34	0.76	0.36	23683
62	0.72	1.66	0.25	1.12	1.83	0.27	0.89	0.41	0.28	46996
63	0.65	1.08	0.18	0.83	1.67	0.22	0.63	0.43	0.33	43143
64	0.39	0.52	0.05	0.70	1.60	0.22	1.61	0.48	0.29	43674
65	0.56	0.03	0.19	0.76	0.82	0.35	2.44	0.61	0.30	21417

66	0.81	0.39	0.21	0.97	0.88	0.26	1.35	0.57	0.40	31783
67	1.85	2.14	0.21	1.90	2.20	0.21	4.58	0.26	0.19	52056
68	1.77	1.63	0.25	2.02	1.77	0.26	1.04	0.36	0.32	50657
69	0.86	1.38	0.15	0.92	1.63	0.16	1.27	0.34	0.36	30691
70	1.01	1.40	0.35	1.35	1.72	0.37	1.20	0.48	0.28	45591

APPENDIX B: 2004 IRT PARAMETERS AND STATISTICS

	Parscale			Xcalibre						
Item	*a*	*b*	*c*	*a*	*b*	*c*	*R*	*P*	r_{pbis}	Omit
1	0.81	1.52	0.19	0.81	1.96	0.18	2.89	0.29	0.25	10793
2	1.20	0.63	0.25	1.19	0.85	0.25	1.16	0.49	0.38	14740
3	0.73	1.86	0.25	1.11	2.37	0.26	1.17	0.31	0.17	30440
4	1.33	1.12	0.18	1.42	1.50	0.19	0.62	0.33	0.39	25344
5	1.01	1.01	0.27	1.28	1.43	0.29	3.91	0.43	0.33	25704
6	0.67	0.88	0.26	0.65	1.30	0.24	1.36	0.48	0.27	20671
7	0.83	0.19	0.26	0.82	0.57	0.26	1.43	0.62	0.30	23916
8	1.23	2.25	0.15	1.74	2.89	0.15	2.42	0.17	0.07	36074
9	0.84	0.70	0.26	0.82	1.05	0.25	0.68	0.50	0.31	19993
10	0.83	0.23	0.33	0.74	0.27	0.28	1.50	0.64	0.27	12424
11	0.89	1.58	0.44	1.64	1.98	0.45	5.17	0.50	0.16	22834
12	0.96	0.43	0.23	1.20	1.62	0.31	3.70	0.50	0.41	53309
13	1.03	0.98	0.20	0.98	1.20	0.19	2.91	0.39	0.34	15097
14	1.29	1.22	0.20	1.42	1.55	0.20	0.90	0.34	0.36	26321
15	1.41	0.88	0.30	1.28	1.09	0.29	0.74	0.48	0.35	16972
16	0.94	0.73	0.23	0.80	0.79	0.18	1.46	0.48	0.32	12539
17	1.41	1.79	0.12	1.79	2.34	0.13	5.03	0.16	0.19	19994
18	1.14	0.49	0.27	1.00	0.65	0.24	1.06	0.55	0.34	18427
19	0.24	2.95	0.13	1.81	2.94	0.31	4.69	0.32	0.07	53160
20	1.08	0.57	0.26	1.01	0.77	0.24	1.30	0.53	0.34	21278
21	1.00	10.62	0.17	2.19	3.00	0.17	2.75	0.17	0.02	40651
22	1.87	1.23	0.21	1.53	1.72	0.21	4.21	0.32	0.34	32290
23	1.20	1.38	0.24	1.12	1.89	0.25	2.54	0.35	0.26	27547
24	0.91	-0.43	0.13	0.90	-0.24	0.11	4.96	0.71	0.27	16250
25	1.40	0.22	0.19	0.20	3.00	0.43	38.79	0.59	0.38	20498
26	1.32	1.79	0.18	1.57	2.37	0.18	3.76	0.23	0.21	41664
27	0.11	4.05	0.16	2.46	3.00	0.42	5.78	0.42	-0.05	34831
28	0.53	1.94	0.17	0.69	3.00	0.20	2.57	0.27	0.16	44958
29	1.00	1.12	0.23	1.17	1.81	0.24	0.67	0.38	0.33	44889
30	1.18	0.03	0.25	1.08	0.39	0.26	1.34	0.66	0.34	26471
31	1.33	1.47	0.14	1.25	2.07	0.14	4.43	0.22	0.30	28768
32	0.96	-0.72	0.27	0.89	-0.80	0.18	3.46	0.80	0.23	8784
33	1.34	1.44	0.17	1.74	1.87	0.17	3.20	0.26	0.33	36203
34	0.86	0.79	0.25	1.29	1.30	0.29	4.06	0.47	0.36	30871
35	1.41	0.77	0.27	1.69	1.39	0.28	0.84	0.48	0.44	49051
36	0.98	1.35	0.21	1.28	1.95	0.22	0.48	0.34	0.31	48196
37	1.07	-0.03	0.22	1.01	0.23	0.21	1.83	0.65	0.34	20662
38	0.61	-0.89	0.00	0.70	-0.24	0.12	5.40	0.71	0.24	17966
39	0.63	-0.58	0.00	0.76	0.82	0.27	3.31	0.64	0.29	36499
40	1.52	0.91	0.16	2.50	3.00	0.25	5.54	0.24	-0.10	53963
41	1.50	1.33	0.27	1.70	1.59	0.27	4.01	0.36	0.29	9004
42	0.13	0.00	0.00	2.50	3.00	0.23	8.89	0.23	-0.07	31279
43	0.35	1.41	0.22	0.72	2.23	0.32	2.49	0.43	0.18	30555
44	0.58	-0.49	0.16	0.67	0.27	0.29	2.64	0.69	0.26	20608
45	0.67	0.39	0.23	0.83	1.25	0.29	1.40	0.55	0.31	39373
46	0.69	1.58	0.30	1.18	2.17	0.32	2.07	0.39	0.20	39709
47	1.45	1.93	0.13	1.67	2.70	0.13	6.39	0.15	0.14	43419
48	0.69	-0.23	0.10	0.64	0.75	0.19	2.47	0.61	0.33	32004
49	1.00	2.21	0.21	1.64	3.00	0.21	1.99	0.22	0.06	51965
50	1.16	0.53	0.22	1.28	1.09	0.25	1.96	0.49	0.43	33320
51	1.41	1.55	0.15	1.53	2.12	0.15	4.52	0.22	0.27	29750
52	0.82	-0.16	0.28	1.02	0.66	0.41	3.65	0.68	0.33	28594
53	1.13	-0.82	0.09	0.84	-0.03	0.27	2.30	0.77	0.31	26486
54	0.93	1.28	0.23	1.20	2.01	0.24	0.42	0.35	0.31	48427
55	1.41	1.03	0.29	1.32	1.57	0.29	1.57	0.42	0.34	31381
56	1.55	0.62	0.17	1.47	0.87	0.16	4.01	0.45	0.48	24440
57	1.01	0.19	0.16	1.10	0.85	0.19	2.48	0.55	0.42	37485
58	1.32	0.68	0.22	1.43	1.08	0.22	1.92	0.46	0.44	32903

59	0.68	1.66	0.17	1.06	2.37	0.17	1.67	0.27	0.28	56167
60	0.62	0.70	0.15	0.85	1.56	0.20	2.85	0.44	0.34	44496
61	0.08	5.71	0.15	2.50	3.00	0.40	5.57	0.40	-0.02	35293
62	1.12	-0.44	0.21	1.23	0.66	0.39	2.79	0.72	0.39	40404
63	0.09	2.70	0.22	2.07	3.00	0.37	3.22	0.37	0.01	46744
64	0.81	-0.18	0.16	1.18	0.92	0.33	2.28	0.63	0.38	41771
65	0.67	0.23	0.11	0.88	1.67	0.22	1.75	0.49	0.42	55927
66	1.25	1.08	0.27	1.62	1.67	0.27	0.99	0.42	0.38	51153
67	1.13	-0.64	0.16	1.16	0.18	0.37	2.55	0.77	0.32	27839
68	0.18	-1.23	0.38	1.78	2.88	0.31	3.47	0.32	0.01	29969
69	1.69	0.61	0.19	1.70	1.27	0.21	0.82	0.46	0.52	49717
70	0.55	0.59	0.16	0.83	1.49	0.23	2.80	0.48	0.33	42655

APPENDIX C: 2003 DIF COMPARISON

		Male			Female			
Item	*a*	*b*	*c*	*a*	*b*	*c*	*χ2*	*p*
1	0.20	1.21	0.47	0.20	1.57	0.47	0.08	0.77
2	1.39	0.80	0.27	1.39	0.63	0.27	77.99	0.00
3	1.06	2.20	0.36	1.06	2.56	0.36	32.27	0.00
4	0.76	1.70	0.41	0.76	1.74	0.41	0.57	0.46
5	0.58	0.62	0.24	0.58	0.18	0.24	43.33	0.00
6	0.78	1.19	0.14	0.78	0.60	0.14	414.23	0.00
7	0.38	-0.65	0.08	0.38	-0.96	0.08	10.55	0.00
8	1.44	0.00	0.14	1.44	2.45	0.14	33.79	0.00
9	0.34	3.32	0.11	0.34	3.03	0.11	8.08	0.01
10	0.86	0.98	0.19	0.86	0.74	0.19	88.13	0.00
11	1.39	2.04	0.16	1.39	2.11	0.16	3.76	0.05
12	0.69	0.34	0.26	0.69	0.06	0.26	30.58	0.00
13	2.28	2.39	0.31	2.28	2.26	0.31	9.87	0.00
14	0.53	0.15	0.08	0.53	-0.08	0.08	13.85	0.00
15	1.32	0.95	0.15	1.32	0.52	0.15	498.65	0.00
16	1.23	-0.52	0.19	1.23	-0.81	0.19	99.44	0.00
17	0.10	0.00	0.00	0.10	0.00	0.00	0.00	1.00
18	0.63	-0.20	0.05	0.63	-0.27	0.05	2.62	0.10
19	0.15	0.00	0.39	0.15	0.00	0.39	0.00	1.00
20	1.57	1.49	0.12	1.57	1.31	0.12	80.12	0.00
21	1.09	0.53	0.21	1.09	0.44	0.21	13.57	0.00
22	1.01	0.95	0.16	1.01	1.23	0.16	150.31	0.00
23	0.48	0.49	0.13	0.48	0.23	0.13	9.59	0.00
24	0.88	-0.20	0.09	0.88	-0.47	0.09	66.49	0.00
25	1.32	0.49	0.19	1.32	0.14	0.19	298.82	0.00
26	0.55	1.34	0.16	0.55	0.86	0.16	100.47	0.00
27	1.26	0.66	0.26	1.26	0.28	0.26	295.64	0.00
28	1.09	-0.56	0.11	1.09	-0.74	0.11	38.66	0.00
29	0.88	0.06	0.21	0.88	-0.23	0.21	56.71	0.00
30	1.12	0.08	0.26	1.12	-0.09	0.26	36.28	0.00
31	0.70	-0.29	0.09	0.70	-0.51	0.09	24.08	0.00
32	1.10	-0.07	0.22	1.10	-0.37	0.22	111.00	0.00
33	0.98	2.07	0.18	0.98	2.29	0.18	23.59	0.00
34	0.96	-0.06	0.14	0.96	-0.28	0.14	56.55	0.00
35	0.10	0.00	0.00	0.10	0.00	0.00	0.00	1.00
36	0.36	3.39	0.26	0.36	2.70	0.26	28.17	0.00
37	1.08	-0.18	0.26	1.08	-0.52	0.26	106.65	0.00
38	0.66	0.69	0.13	0.66	0.28	0.13	102.86	0.00
39	1.32	1.16	0.17	1.32	0.81	0.17	294.42	0.00
40	0.60	2.04	0.14	0.60	2.09	0.14	1.16	0.28
41	1.13	0.30	0.15	1.13	0.04	0.15	141.19	0.00
42	0.98	2.70	0.18	0.98	2.36	0.18	27.51	0.00
43	1.00	0.95	0.19	1.00	0.71	0.19	83.78	0.00
44	0.04	0.00	0.69	0.04	0.00	0.69	0.00	1.00
45	0.75	0.77	0.31	0.75	0.54	0.31	32.28	0.00
46	0.49	0.61	0.10	0.49	0.35	0.10	13.53	0.00
47	0.17	-2.18	0.77	0.17	-2.45	0.77	0.03	0.84
48	0.17	2.01	0.21	0.17	1.15	0.21	1.41	0.23
49	1.04	-0.03	0.24	1.04	-0.43	0.24	155.81	0.00
50	0.70	2.03	0.25	0.70	1.66	0.25	72.70	0.00
51	0.87	0.77	0.30	0.87	0.37	0.30	144.65	0.00

52	0.57	-0.12	0.07	0.57	-0.40	0.07	25.05	0.00
53	0.88	-0.44	0.13	0.88	-0.42	0.13	0.17	0.68
54	0.11	2.36	0.22	0.11	34.17	0.22	7.13	0.01
55	0.03	-1.57	0.75	0.03	0.00	0.75	0.00	0.91
56	0.10	2.49	0.18	0.10	0.00	0.18	0.90	0.35
57	0.91	1.90	0.18	0.91	1.86	0.18	0.82	0.37
58	2.05	1.31	0.19	2.05	1.01	0.19	200.16	0.00
59	0.87	-0.01	0.07	0.87	-0.27	0.07	73.00	0.00
60	0.51	2.48	0.19	0.51	2.33	0.19	4.11	0.04
61	0.61	0.87	0.27	0.61	0.38	0.27	7.36	0.01
62	0.65	1.71	0.24	0.65	1.45	0.24	30.60	0.00
63	0.62	1.19	0.18	0.62	0.86	0.18	50.32	0.00
64	0.40	0.84	0.10	0.40	0.48	0.10	12.85	0.00
65	0.52	-0.12	0.17	0.52	-0.16	0.17	0.24	0.63
66	0.74	0.41	0.20	0.74	0.17	0.20	29.88	0.00
67	1.78	2.02	0.21	1.78	2.03	0.21	0.05	0.82
68	1.66	1.58	0.25	1.66	1.48	0.25	15.23	0.00
69	0.80	1.52	0.15	0.80	1.17	0.15	129.07	0.00
70	0.89	1.45	0.35	0.89	1.20	0.35	40.44	0.00

APPENDIX D: 2004 DIF COMPARISON

		Male			Female			
Item	*a*	*b*	*c*	*a*	*b*	*c*	*χ2*	*p*
1	0.86	1.83	0.18	0.86	1.42	0.18	163.32	0.00
2	1.23	0.68	0.24	1.23	0.25	0.24	425.86	0.00
3	0.92	2.45	0.25	0.92	1.36	0.25	671.50	0.00
4	1.36	1.35	0.18	1.36	0.81	0.18	605.82	0.00
5	1.08	1.19	0.27	1.08	0.69	0.27	367.34	0.00
6	0.67	0.92	0.25	0.67	0.57	0.25	97.73	0.00
7	0.81	0.07	0.25	0.81	-0.17	0.25	41.84	0.00
8	0.06	0.00	0.56	0.06	0.00	0.56	0.00	0.95
9	0.88	0.74	0.26	0.88	0.55	0.26	46.30	0.00
10	0.83	0.17	0.33	0.83	-0.03	0.33	32.75	0.00
11	1.10	1.84	0.44	1.10	1.33	0.44	166.35	0.00
12	0.96	0.44	0.23	0.96	0.16	0.23	67.29	0.00
13	1.06	1.16	0.20	1.06	0.67	0.20	494.99	0.00
14	1.44	1.51	0.20	1.44	0.80	0.20	1026.83	0.00
15	1.38	1.03	0.29	1.38	0.30	0.29	1251.40	0.00
16	0.89	0.77	0.21	0.89	0.15	0.21	621.32	0.00
17	5.53	4.18	0.16	5.53	2.43	0.16	0.14	0.71
18	1.06	0.53	0.24	1.06	-0.20	0.24	954.85	0.00
19	0.21	3.40	0.12	0.21	3.29	0.12	0.16	0.69
20	0.99	0.57	0.23	0.99	-0.19	0.23	898.99	0.00
21	0.02	6.96	0.80	0.02	0.00	0.80	0.41	0.53
22	1.80	1.52	0.21	1.80	0.92	0.21	753.09	0.00
23	1.26	1.72	0.24	1.26	1.09	0.24	562.49	0.00
24	0.90	-0.41	0.22	0.90	0.00	0.22	173.08	0.00
25	0.99	0.15	0.12	0.99	-1.36	0.12	3579.48	0.00
26	1.53	2.23	0.18	1.53	1.56	0.18	391.12	0.00
27	0.15	10.32	0.39	0.15	0.20	0.39	132.30	0.00
28	0.24	2.88	0.05	0.24	-0.11	0.05	656.77	0.00
29	1.06	1.45	0.22	1.06	0.68	0.22	826.47	0.00
30	1.08	-0.07	0.22	1.08	-0.56	0.22	299.90	0.00
31	1.39	1.89	0.14	1.39	1.14	0.14	855.30	0.00
32	0.22	0.00	0.00	0.22	0.00	0.00	0.00	1.00
33	1.62	1.81	0.17	1.62	1.12	0.17	773.73	0.00
34	0.95	0.93	0.25	0.95	0.43	0.25	325.51	0.00
35	1.33	0.91	0.26	1.33	0.35	0.26	494.66	0.00
36	0.99	1.63	0.21	0.99	1.23	0.21	143.94	0.00
37	0.87	-0.24	0.16	0.87	-1.19	0.16	646.65	0.00
38	0.60	0.00	0.00	0.60	0.00	0.00	0.00	1.00
39	0.61	-0.71	0.03	0.61	-0.62	0.03	7.98	0.01
40	0.01	2.78	0.88	0.01	1.51	0.88	0.00	0.92
41	1.84	1.63	0.27	1.84	0.94	0.27	961.17	0.00
42	0.03	2.95	0.79	0.03	0.00	0.79	0.84	0.36
43	0.46	1.77	0.25	0.46	1.28	0.25	58.11	0.00
44	0.66	-0.32	0.27	0.66	0.00	0.27	38.75	0.00

45	0.70	0.49	0.27	0.70	1.14	0.27	221.48	0.00
46	0.38	1.92	0.21	0.38	-0.10	0.21	642.07	0.00
47	1.45	2.73	0.13	1.45	1.73	0.13	441.81	0.00
48	0.71	-0.12	0.22	0.71	0.67	0.22	419.97	0.00
49	1.35	2.79	0.21	1.35	2.31	0.21	25.89	0.00
50	1.16	0.59	0.22	1.16	0.15	0.22	341.13	0.00
51	1.51	1.97	0.15	1.51	1.26	0.15	671.20	0.00
52	0.89	-0.18	0.31	0.89	-0.40	0.31	29.74	0.00
53	1.16	-0.63	0.34	1.16	0.00	0.34	352.97	0.00
54	0.95	1.56	0.22	0.95	1.09	0.22	206.96	0.00
55	1.43	1.26	0.28	1.43	0.67	0.28	605.29	0.00
56	1.57	0.74	0.17	1.57	0.14	0.17	1053.16	0.00
57	1.01	0.19	0.15	1.01	0.02	0.15	41.57	0.00
58	1.33	0.81	0.21	1.33	0.21	0.21	792.23	0.00
59	0.71	2.13	0.16	0.71	1.39	0.16	286.81	0.00
60	0.57	0.72	0.11	0.57	0.24	0.11	108.05	0.00
61	2.10	7.11	0.40	2.10	7.77	0.40	0.00	0.95
62	1.10	-0.62	0.19	1.10	-0.91	0.19	67.64	0.00
63	0.05	0.00	0.35	0.05	0.00	0.35	0.00	1.00
64	0.86	-0.22	0.19	0.86	-0.02	0.19	34.54	0.00
65	0.66	0.16	0.09	0.66	-0.16	0.09	43.98	0.00
66	1.36	1.36	0.27	1.36	0.64	0.27	775.56	0.00
67	1.13	-0.80	0.18	1.13	-0.83	0.18	1.02	0.31
68	0.19	-2.24	0.55	0.19	0.00	0.55	3.09	0.08
69	1.64	0.67	0.19	1.64	0.26	0.19	352.15	0.00
70	0.53	0.56	0.12	0.53	0.15	0.12	58.05	0.00

REFERENCES

[1] Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*; (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

[2] Camilli, G. and Shepard, L. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage.

[3] Embretson, S. E., and Reise, S. P. (2000). Item response theory for psychologists. London: Lawrence Erlbaum Associates, Publishers.

[4] Farhady, H. & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, 29, 132–141. doi:10.1017/S0267190509090114.

[5] Hambleton, R. K., Swaminathan, H., & Rogea, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

[6] Hambleton RK. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, *6* (3), 535-556.

[7] Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

[8] Henning, G. (1987). A guide to language testing: development, evaluation, research. Cambridge: Newbury House Publishers.

[9] Lord, F. M. (1980). Applications of item response theory to practical problems. Hillsdale, NJ: Erlbaum.

[10] Pae, T. & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language testing*, 23(4), 475-96.

[11] Raju. N. S. (1988). The area between two item characteristic curves. *Psvchometrika 53,495-502.*

[12] Reeve, B. B. (2003). An introduction to modern measurement theory. Retrieved from http://appliedresearch.cancer.gov/areas/cognitive /immt.pdf. on 6 July 2006.

[13] Smith, R. M. (1991). IPARM: item and person analysis with the RASCH model. Chicago, IL: Mesa Press.

[14] Teresi J. A, Kleinman M, and Ocepek-Welikson K. (2000). Modern psychometric methods for detection of differential item functioning: Application to Cognitive Assessment Measures. *Statistics in Medicine,* 19, 1651-83.

[15] Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group threshold differences in trace lines. In Wainer, H. and Braun, H., editors. *Test validity*. Hillsdale, NJ: Lawrence Erlbaum, 147-69.

[16] Zumbo, B. D. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going? *Language assessment quarterly*, 4 (2), 223-33.

[17] Zumbo, B. D. and Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In D. Kaplan (Ed.), *the SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: SAGE.

Alireza Ahmadi received his Ph.D. in TEFL from the University of Isfahan in 2008, his MA in TEFL from Shiraz University in 2002 and his BA in English Translation from the University of Allameh Tabatabaei in 2000.

Currently, he is an assistant professor in the Department of Foreign Languages and Linguistics at Shiraz University, Iran. His main interests include Language Assessment and Second Language Acquisition.

Nathan A. Thompson has a Ph.D. in Psychometric Methods from the University of Minnesota. He is currently the Vice President of Assessment Systems Corporation and adjunct faculty at the University of Cincinnati. He is primarily interested in Computerized Adaptive Testing and the development of test software tools.