広島文教女子大学紀要　47，2012

【原著】

Differential Item and Test Functioning Across Academic Disciplines

Judith Runnels

差異項目機能（DIF）と差異試験機能（DTF）の学科別測定

Judith Runnels

Abstract

Differential item functioning (DIF) analyses are used to determine if there are any items that affect the probability of particular groups of test-takers endorsing that item, after controls for ability are taken into account. If DIF occurs on a wide-scale, this means that test scores do not represent the same measurement over the population of test-takers. This is known as differential test functioning (DTF). This study examined the item functioning of an in-house designed low-stakes achievement vocabulary test designed to measure how well second-year students in 4 different academic disciplines acquired words on a 250 word English study list. The same test has been previously examined using Rasch analysis, with the purposes of making validity arguments and highlighting items that were causing unexpected response patterns (Runnels, 2011). The current analysis offers additional validity evidence related to score equivalence across majors. It was found that even though DTF is unlikely, there were several items that favored and hindered some majors. The importance of establishing a process to check for DTF and DIF, especially when the test-takers are from different disciplines of study and even for low-stakes tests, is discussed.

Keywords: Rasch Model, differential item functioning, differential test functioning, academic discipline

Introduction

A major concern in test development is that the interpretations and ultimate uses based on test scores are valid and reasonable (Messick, 1989; Bachman, 1990). It is also of the utmost importance that neither the test as a whole nor any individual item on the test is favored by any characteristic of a test-taker (Lumley & O'Sullivan, 2005). Test items are considered biased if characteristics other than those being measured change the

probability that a person will get an item correct (Lord & Novick, 1968). Examining this internal bias requires a statistical approach to assessment evaluation that is capable of detecting the existence of some inherent characteristic in the test that is not only causing groups of test takers of the same ability levels to respond differently but also to detect the source of this internal bias (Geranpayeh & Kunnan, 2006). This is known as differential item functioning (DIF) or item bias (most authors now use DIF rather than item bias although they share the same definition- Angoff, 1993; Cole, 1993). DIF occurs when examinees with the same level of ability from two different groups have different probabilities of endorsing an item (Clauser & Mazor, 1998) and can result in higher or lower scores for test takers within that group (Swaminathan & Rogers, 1990; Chang & Mazzeo (1994); Donoghue, Holland & Thayer, 1993). Differential test functioning (DTF) is when the total score functions differently across groups such that the final scores do not represent the same measurement across the population of test-takers (Raju, van der Linde & Fleer, 1995).

Karami (2011) poses that the statistical procedure used to identify any item bias could be considered equivalent with statistical bias, which may entail an over or underestimated parameter in a model because of lack of a control group, for one (Camilli, 2006; Wiberg, 2007). This is a common criticism of using traditional deterministic statistics for DIF analyses: some items will either falsely exhibit DIF or exhibit none at all since there is no adequate control for ability across all members of groups (Karami, 2011). Methods such as logistic regression are, nonetheless, commonly employed in DIF analyses. Bruckner *et al*. (2007) used logistic regression to detect differences in sub-groups of test-takers' responses to vocabulary multiple-choice questions (MCQ). Runnels (2011) checked for DIF across groups using ANOVAs on an MCQ vocabulary test and found no significant differences. As Elder (1996) notes however, test bias investigations are often concerned more with the nature rather than the magnitude of differences between groups, which makes the Rasch model particularly suitable for these types of analyses. Furthermore, the specific procedure for equating abilities across groups using the Rasch model is well established (Wolfe, 2004).

Rasch-based methods analyze the full spectrum of abilities and not only a proportion of high and low ability test-takers or the overall raw score of the group. The Rasch model is prescriptive rather than deterministic, uses probability to determine the relationship between a raw score and a person's ability on an item-by-item basis (Bond & Fox, 2001). Rasch takes into account both item difficulty and person ability (Rasch, 1980). Rasch analysis converts a participant's raw test score into a ratio of success to failure and then into the logarithmic odds that the person will respond correctly to an item (a logit; Smith, 2000). All logits are plotted on a single scale and used as an estimate of ability for a test-taker and difficulty of an item. The relationship between these two probabilities is known as the Rasch Simple Logistic Model (Wright & Stone, 1979) and has the capability of identifying people or items that exhibit unexpected response patterns. For DIF in

particular, the Rasch model identifies items for which being a member of a certain group of test-takers affects the probability of successfully responding to that item compared to other groups.

Most commonly, Rasch is used to measure differences in gender, ethnicity and native language on pedagogical performance tests (Meade & Fetzer, 2009). In terms of language, one of the first publications in DIF using the Rasch model was by Chen and Henning (1985) who examined bias in a university placement test for Chinese and Spanish test takers. Elder (1996) also performed a study using Rasch to analyze whether language background lead to DIF on a test of English. Takala and Kaftandjieva (2000) investigated DIF using a modified Rasch model on the vocabulary section of Finnish Foreign Language Certificate Examination, which is a high stakes exam the Finnish government requires all high school students complete. They found that over a quarter of items showed DIF according to gender although concluded that overall the test was not biased since the number of items exhibiting female favored DIF equaled the number of items exhibiting male favored DIF. Nonetheless, they note that any time there are differences in test-takers, DIF analyses are important, especially when item banks are being employed.

More recently, DIF analysis has been used in education to measure differences across academic disciplines of study (Alavi, Rezaee & Amirian, 2011). In one of the first such studies, Alderson and Urquhart (1985) studied test-takers with the same first language, using academic major of study to define their groups. They analyzed the performance of students from different disciplines on an English-for-specific-purposes reading test and found that indeed academic discipline affected test performance, although the effects were admittedly inconsistent. Pae (2004) performed DIF analyses on the National Korean University and College entrance exam to determine that nearly thirty percent of items favored either a science or a humanities major. Alavi, Rezaee & Amirian (2011) studied DIF in the University of Tehran English Proficiency Test for humanities and science and engineering groups and found that unintentionally, there were items that favored one and hindered another major.

In the current analysis, it is important to consider the possibility of DIF since the test takers consist of non-English majors and therefore, verification that the test is not biased towards one major or another is necessary. The current study was designed to investigate the DIF and DTF of a second-year vocabulary test using Rasch analysis. A detailed Rasch analysis was previously performed by Runnels (2011) who concluded that the test was not challenging most test-takers: it was not effectively determining how well students had achieved acquiring vocabulary words from a 250 word study list. Nonetheless, employing the Rasch model for a DIF analysis was deemed appropriate since the construct of the instrument was previously examined (Wright & Mok, 2004). No significant differences across majors were previously found, but since this was based on raw-scores only, this investigation lacked any specific comparisons across majors for test-takers of the same

ability. One of the major assumptions made by Runnels (2011) was that all test-takers were equivalent, despite being members of 4 departments of different disciplines of study (Early Childhood Education, Welfare, Psychology and Nutrition), since they all participated in the same English classes. Since some of the curriculum may have favored some disciplines over others (for example, the food and health unit may have favored the Nutrition department), it is important to ensure that the test is taking equivalent measurements across groups and not simply individuals (Camilli, 2006). It is expected that some items will exhibit differential item functioning and that subsequent measures, either in the form of grade adjustments or test modifications will be required.

Method

Participants

The test-takers were 294 female second year non-English majors from Hiroshima Bunkyo Women's University (aged 20 and 21 years old), a private university in Hiroshima City, Japan. They were in one of 11 different classes organized according to major (Early Childhood Education, Welfare, Psychology and Nutrition). Japanese (L1) and English (L2) are respectively the first and second languages of all test-takers.

Instrument

The test (administered in 2011) consisted of 83 multiple choice questions designed to determine how well students had acquired words from a 250 word study list related to the curriculum they studied in two periods of 90 minutes every week. Nine percent of words were specific to the lesson content of the curriculum while the remaining 91% fell within the 3000 most frequent words of English (Nation, 2001). There were 4 question types: L1-L2 translations (14 items) for single words and within sentences, L2-L1 (24 items) for single words and within sentences, sentence completion in L2 (34 items) and matching an object or activity in a picture with its word or phrase (11 items) (Nation, 2001). Details about the validity and reliability of the test can be found in Runnels (2011), who demonstrated no significant differences across majors using ANOVAs. The mean score of the test was reported at 86.2% (SD = 8.7%). Runnels' (2011) results also indicated that all items fit well within the Rasch model since the infit statistics (in the form of a mean-standardized score, or MNSQ) for all items were between 0.7 and 1.3, an acceptable spread for a low-stakes test (Bond & Fox, 2007). Three items (46, 49, 71) fell outside the acceptable range for outfit (measured in the form of a z-standardized score or ZSTD), all with statistically significant ZSTDs (see Linacre, 2007).

Procedure

The test was taken in the students' usual class time (up to 90 minutes) and in their regular classrooms. The test was administered using www.classmarker.com©, an online testing site. Classmarker.com© both forces a selection before the test-taker may proceed to the following question as well as randomizes distracter order. WINSTEPS® Rasch software

Version 3.72.4 (Linacre, 2008) was used to analyze the results of the test and produced both the item strata and the differential item analyses outputs.

Item separation strata are used to identify statistically distinct difficulty levels according to the responses to all items (Wright & Masters, 2002). They demonstrate the range of item-difficulties that have been included. Item strata are calculated using the following formula (Beglar, 2010):

$$\text{Item strata} = (4Gitem + 1/3)$$

where *Gitem* is the Rasch item separation value (derived by dividing the item standard deviations by the average measurement error). Smith (2001) requires a minimum two level difficulty level in order to deem the measurement tool representative of the assessed content.

DIF was examined for each item on the vocabulary test. WINSTEPS® has several ways of testing for DIF. The first hypothesis is that the item has the same difficulty for two groups, the second is the hypothesis that the item has the same difficulty as the average difficulty for all groups, and the third hypothesis is that the item has no overall DIF across all groups (Linacre, 2004). For the current analysis, the third hypothesis will be checked first to determine which items are exhibiting DIF across all majors. If there are no items with any evident DIF, the second hypothesis will be checked to determine if any items have a statistically significant differing difficulty to the average difficulty across majors. Finally, if necessary, items which have different difficulties across two majors will also be measured (the first hypothesis) to determine where these differences specifically lie. Altogether, testing these three hypotheses also reveals if any DIT is occurring.

The former two of these hypotheses employ a t-test approach proposed first by Wright and Stone (1979) to determine the amount of DIF (Smith, 2004). WINSTEPS® (Linacre, 2010) uses the following formula:

$$t = \frac{d_{i2} - d_{i1}}{\sqrt{(s^2_{I2} - s^2_{I1})}}$$

where d_{i2} is the difficulty of item *I* in calibration based on two groups, d_{i1} is the difficulty of item *i* in calibration *l*, s^2_{I1} is the standard error of the estimate for d_{i1} and s^2_{I2} is the standard error of the estimate for d_{i2} (Alavi & Karami, 2010). Instead of fit statistics being calculated for each individual item, for DIF tests, Rasch produces the between-class mean squares as the Between-Group fit statistics (see Smith & Plackner, 2009). Like the individual item MNSQ, the between-group fit indices also test the hypothesis that the dispersion of the group measures fits with the expectations of the Rasch model. The mean-square is the chi-square value (of the previously described formula) divided by its degrees

of freedom and it represents the size of the misfit to the Rasch model. Acceptable values for MNSQs are between 0.7-1.3 and from -2.0 to 2.0 for ZSTDs on a low stakes test (Bond & Fox, 2007).

In addition, the Mantel-Haenszel statistic will be calculated. This is a common investigation tool for DIF which employs log-odds estimators (Mantel & Haenszel, 1959). The sample is divided into different classification groups and then sliced into strata by ability - it is said to be equivalent to raw score usage (see Rogers & Swaminathan, 1993). Linacre (2007) notes that the Mantel-Haenszel and t-tests used in WINSTEPS® should produce similar results, due to the fact that they are based on the same logit-linear theory. Both will be reported herein.

Finally, Alavi and Karami (2010) postulate that a major problem with using significance tests in DIF analyses is that they are sensitive to sample size. If the sample sizes are large, minor differences will show up as significant while the opposite occurs for small sample sizes: large differences between the groups may not show up as significant. In the present study, the sample size can reasonably be considered as large and therefore, using the significance tests directly from the software output, may not be justified. Linacre (2007) also notes that for the first hypothesis in particular (the comparisons across two majors) that DIF tests can be doubtful in the context of Rasch analysis. Despite statistically significant differences, the impact of one item may have too little an influence on the meaning or eventual practical use of the test results, and therefore both statistical significance and a large enough sized difference should be required before any action regarding potential item biases is taken. For the latter, a logit difference of at least 0.5 is required although this may differ depending on the stakes of the test (Linacre, 2007). The former will be addressed by employing a Bonferroni correction test. This entails distributing the alpha level (in this case, 0.05) across all of the comparisons such that 0.05 is the sum of the alpha levels (see Thompson, 2006). By dividing the alpha level by the number of items, this produces a new level at which to judge significance.

Results

Summary Statistics

Summary statistics for the vocabulary test are shown in Table 1. These include the mean measure of items, the model error as well as infit and outfit MNSQs and ZSTDs. The mean MNSQ for both infit and outfit falls within the acceptable range for a low-stakes test (Bond & Fox, 2007). Cronbach's alpha for items is 0.95 and the separation strata for non-extreme items (items for which there was not a 0% or a 100% success rate) was 4.42: an adequate level of distinct difficulties that demonstrates that the scale discriminated effectively between test-takers (Smith, 2001).

Table 1.
Summary statistics.*

	SCORE	INFIT		OUTFIT	
		MNSQ	ZSTD	MNSQ	ZSTD
MEAN	86.2	.99	.2	.88	-.1
S.D.	8.7	.09	1.2	.44	1.4
SEPARATION	4.42	Item	RELIABILITY	.95	

* MNSQ − Mean-square, ZSTD − standardized z-score.

Differential Item Analyses

The results of the hypothesis test that items have no overall DIF across all groups are shown in Table 2. There are 8 items with statistically significant DIF at the $p<0.05$ level (Items 17, 28, 52, 57, 65, 74, 79, 80). Of these 8 items, 6 of them are exhibiting DIF across all majors, while 2 of them (items 17 and 79) are exhibiting DIF for 2 majors. The Bonferroni correction test puts the new significance level at 0.0006, at which only 1 out of the 8 items shows significant DIF (item 65).

Table 2.
*Items with DIF at the p<0.05 level**

Person	SUMMARY DIF			BETWEEN-CLASS		Item
CLASSES	CHI-SQUARE	D.F.	PROB.	MEAN-SQUARE	t=ZSTD	Number
2	4.8828	1	.0271	2.9108	1.3789	17
4	8.0885	3	.0439	.7865	-.0104	28
4	9.5431	3	.0227	.6079	-.2896	52
4	10.7915	3	.0128	1.1145	.4074	57
4	18.4609	3	.0003	1.5126	.8156	65
4	9.4414	3	.0237	.6822	-.1675	74
2	4.4745	1	.0344	2.0673	1.0524	79
4	9.4464	3	.0237	.9620	.2251	80

* The first column, Person CLASSES is the count of groups (in this case, major) with estimable DIF for the item. The SUMMARY DIF CHI-SQUARE column is the sum of the squared normalized t-statistic value for each item. D.F. represents the degrees of freedom or the count of majors minus 1 contributing to the chi-square. The PROB. is the probability of the chi-square (any value under 0.05 is statistically significant) followed by the BETWEEN-CLASS MEAN-SQUARE. The sixth column, the t=ZSTD represents the significance of the MEAN-SQUARE standardized as a unit-normal deviate (t-statistic with infinite degrees of freedom). The final column, the item number, accords with the question number on the test.

For the hypothesis that the item has the same difficulty as its average difficulty for all groups, statistically significant differences at an alpha level of 0.05 are shown in Table 3 (a total of 11 items as follows: 10, 19, 20, 28, 52, 57, 59, 65, 74, 80). Given that these items span

Table 3.

*Item difficulty by major compared to average difficulty for all groups**

Person CLASS	OBSERVATIONS COUNT	AVERAGE	DIF SCORE	DIF MEASURE	DIF SIZE	DIF S.E.	DIF t	Prob.	Item Number
B	44	.86	-.08	.79	1.09	.47	2.33	.0248	10
A	128	.98	-.02	-.81	1.28	.59	2.16	.0324	19
C	39	.95	-.04	-.92	1.59	.75	2.11	.0416	20
A	128	.91	-.04	.73	.71	.32	2.25	.0263	28
C	39	.33	-.22	3.31	1.09	.37	2.92	.0060	52
A	128	.76	-.08	2.01	.57	.22	2.54	.0123	57
B	44	.84	.15	1.00	-1.05	.44	-2.38	.0222	59
A	128	.59	-.08	2.97	.42	.20	2.09	.0384	65
C	39	.74	.26	1.19	-1.37	.40	-3.42	.0015	65
B	44	.66	-.16	2.20	1.02	.35	2.87	.0064	74
D	81	.81	-.07	1.03	.62	.31	1.99	.0498	80

* The Person CLASS column represents the test-takers' majors (Early Childhood Education (A), Welfare (B), Psychology (C), and Nutrition (D)). The OBSERVATIONS column is what was seen in the data: COUNT is the number of observations of the classification used for DIF estimation and AVERAGE is the average observation on the classification where COUNT* AVERAGE equals the total score of person class on the item (Linacre, 2007). Following this is the DIF SCORE, which represents the difference between the observed and the expected average observations. The DIF MEASURE is the item difficulty for this class. The DIF SIZE column indicates the difference between the DIF MEASURE for this class and the baseline difficulty where a positive value shows that the item is more difficult than expected for that major, and a negative value shows that that item is favoring the according major. DIF S.E. indicates the standard error of the difference and the DIF t is the student's t-statistic. Finally, PROB. is the probability of the t-value and the last column corresponds to the item number on the test. As can be seen in Table 3, there are 12 items across the 4 majors that have statistically significant different difficulties. Despite this, applying the Bonferroni correction test to this analysis reduces the alpha level to 0.0006. At this level of significance, there are no items that are significantly different across groups.

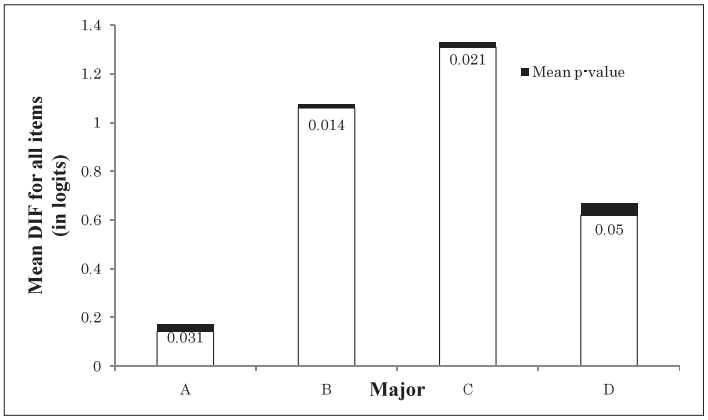

Figure 1. The mean DIF levels for all majors in logits where Major A is Early Childhood Education, Major B is Welfare, Major C is Psychology and Major D is Nutrition.

the four majors, the mean DIF and p-value for each major was calculated and is shown in Figure 1. It can be seen that while Major A (Early Childhood Education) exhibited the highest number of items with DIF (4 in total), taking the mean DIF for all of these items, reduced the DIF to near zero logits. Major C (Psychology) is overall the major with the greatest mean DIF. Linacre (2007) requires at least a 0.5 logit difference in order for DIF to be a concern, and while there is at least a 0.5 overall logit difference between Major C (Psychology) and Major D (Nutrition) and Major A (Early Childhood Education) as well as between Major B (Welfare) and Major A (Early Childhood Education), these are not significant differences, eliminating the possibility of DTF.

A final check is performed to determine specifically which items exhibit differences between majors. Table 4 displays part of the output for the hypothesis that the item has the same difficulty across two groups (statistically significant differences only are provided). In total, there are 17 items exhibiting significant differences (3, 10, 21, 28, 33, 41, 46, 50, 51, 52, 53, 57, 59, 65, 70, 74, 80). Following a Bonferroni test, of these 17 items, only 1 of them is significant at an alpha level of $p<0.0006$ (item 65). It should be noted that some of the items that show up as significant on a Rasch analysis, do not when using Mantel-Haenszel criteria. Of these 17 items, only 7 of them have significant Mantel-Haenszel statistics at $p<0.05$ (items 41, 52, 59, 65, 70, 74, 80). Consulting the raw data, an additional two items exhibited significantly different Mantel-Haenszel statistics: items 7 (chi-square = 4.318, p = 0.383) and 72 (chi-square = 4.290, p = 0.393).

Table 4.
*Item score significant differences across major**

Person CLASS	DIF MEASURE	DIF S.E.	Person CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	Welch t	d.f.	Prob.	Mantel-Haenszel Chi-squ	Prob.	Size CUMLOR	Item Number
A	-1.95	1.01	D	.48	.36	-2.42	1.07	-2.26	194	.0248	3.6005	.0578	-1.86	3
A	.11	.40	C	1.34	.39	-1.24	.56	-2.22	118	.0286	1.8270	.1765	-1.23	21
A	.11	.40	D	1.12	.30	-1.01	.50	-2.02	205	.0447	2.1085	.1465	-1.28	21
A	.73	.32	C	-.92	.75	1.65	.82	2.01	72	.0477	1.8560	.1731	1.83	28
A	1.57	.25	D	2.49	.25	-.91	.35	-2.60	195	.0100	6.5632	.0104	-1.05	41
A	3.13	.20	C	2.16	.36	.97	.41	2.38	84	.0194	1.2672	.2603	.56	46
A	3.01	.20	B	2.20	.35	.81	.41	1.99	94	.0500	3.1738	.0748	.94	50
A	2.06	.22	C	3.31	.37	-1.25	.43	-2.88	87	.0050	4.1754	.0410	-1.07	52
A	2.01	.22	B	1.00	.44	1.01	.50	2.04	89	.0442	.9464	.3306	.67	57
A	2.01	.22	C	.65	.45	1.36	.50	2.70	79	.0084	1.8386	.1751	.88	57
A	2.01	.22	D	1.21	.30	.80	.37	2.16	181	.0321	2.1524	.1423	.69	57
A	2.25	.21	B	1.00	.44	1.25	.49	2.55	87	.0126	3.7479	.0529	1.34	59
A	2.97	.20	B	1.95	.37	1.03	.42	2.46	93	.0155	2.5114	.1130	.79	65
A	2.97	.20	C	1.19	.40	1.78	.45	4.00	79	.0001	8.3775	.0038	1.73	65
A	2.73	.20	C	1.90	.36	.83	.41	2.00	84	.0489	.0500	.8231	.22	70
A	.83	.31	B	2.20	.35	-1.37	.47	-2.93	119	.0041	8.0769	.0045	-1.60	74
A	-.50	.52	C	.85	.43	-1.35	.67	-2.01	128	.0465	.7764	.3782	-1.49	80
A	-.50	.52	D	1.03	.31	-1.53	.60	-2.54	206	.0117	4.4186	.0355	-1.43	80

B	.79	.47	D	-.62	.53	1.41	.71	2.00	115	.0483	1.3409	.2469	1.23	10
B	2.57	.34	C	1.49	.38	1.08	.51	2.10	80	.0386	1.9178	.1661	1.15	33
B	1.81	.37	C	3.31	.37	-1.50	.53	-2.85	80	.0055	4.4500	.0349	-1.79	52
B	1.00	.44	C	2.41	.35	-1.41	.57	-2.49	80	.0150	3.6385	.0565	-2.12	59
B	1.00	.44	D	2.05	.26	-1.05	.51	-2.05	91	.0431	4.2082	.0402	-1.53	59
B	1.95	.37	D	2.85	.25	-.91	.44	-2.06	96	.0422	1.1096	.2922	-.56	65
B	2.92	.34	C	1.90	.36	1.02	.50	2.05	80	.0436	4.2490	.0393	2.18	70
B	2.20	.35	D	.93	.32	1.27	.48	2.68	107	.0085	3.6141	.0573	1.09	74
C	2.79	.36	D	1.90	.26	.88	.44	1.99	91	.0491	2.3102	.1285	.85	51
C	3.31	.37	D	2.10	.26	1.21	.45	2.67	88	.0090	1.6342	.2011	.70	52
C	1.19	.40	A	2.97	.20	-1.78	.45	-4.00	79	.0001	8.3775	.0038	-1.73	65
D	.83	.33	A	-.26	.47	1.09	.57	1.92	206	.0565	.6418	.4231	.73	53
D	2.85	.25	C	1.19	.40	1.66	.47	3.55	84	.0006	2.4313	.1189	.80	65

* The DIF CONTRAST measure illustrates the difference between the DIF MEASUREs, or the difference in difficulty of the item between the two groups (a difference of at least 0.5 logits is required for the DIF to be noticeable). A positive DIF contrast represents that the item was more difficult for the CLASS (major) that is listed on the left whereas a negative DIF contrast demonstrates that the item was easier for the major on the left. The PROB illustrates the probability of observing this amount of contrast by chance. The JOINT S.E. is the standard error of the DIF CONTRAST, the Welch t gives the DIF significance as a t-statistic.

The Mantel-Haenzel chi-square statistic is also included. In this case, the Prob. is the probability of observing these data when there is no DIF, based on a chi-square value with 1d.f. Finally, the Size CUMLOR column shows the cumulative log-odds ratio in logits and is used as an estimate of the size and direction of the DIF.

Summarizing all of the tests herein, there are a total of 23 items displaying DIF of some kind, whether that be due to the three hypotheses tests from WINSTEPS®, or the Mantel-Haenszel statistic. Out of the 23 items, 13 of them were highlighted by one of the hypothesis tests, 3 of them by two of the tests. Moving forward, it is the items that got flagged by 3 or 4 of the tests that may require revisiting (items 28, 52, 57, 59, 65, 74 and 80) although only 1 item (65) was consistently flagged after making the required significance level more severe.

Discussion

The results of the current study indicate that there are a potential 7 items with significant DIF in the vocabulary test. One problem with the analyses is the usage of significance tests as criteria for identification of DIF (Alavi & Karami, 2010). In the current analysis, this issue was addressed by employing the Bonferroni correction test which revealed that only one item was likely causing DIF (item 65). Item 65 is therefore certainly an item that ought to be modified: Early Childhood Education and Nutrition majors found it significantly more difficult compared to Welfare and Psychology majors. On the other hand, Item 52 was easier for Early Childhood Education majors and Welfare students than for Psychology majors. Item 57 was more difficult for Early Childhood Education students than all of the other majors, but they found item 80 easier. Overall, the DTF evidence is not strong enough to dismiss using this test on the grounds of internal bias, although

future administrations of it could certainly benefit from some items being explored and modified, particularly item 65 and possibly the 7 items that were highlighted by three quarters or all of the tests employed in the analysis. The test did in fact, seem to hold up reasonably well under DIF scrutiny, so that for the most part major concern does not need to be given to performance across academic discipline.

Nonetheless, the reason as to why the DIF exists for this test is only speculative. For item 65 in particular, which was more difficult for Early Childhood Education and Nutrition majors, than for Welfare and Psychology, the question was as follows:

Questions 65 and 66 are a set. Choose the word that best fits the blank for number 1.
A: Hi Amy, you don't look so good. What's wrong?
B: Hey Elizabeth! I feel (1)_____. I have a sore throat and a fever.
A: Oh no! You should go to the _____ and get some rest.

a) Terrible (correct answer)
b) Worry
c) Anything
d) Cool

This item had a 60% success rate with 6% selecting 'worry' and 33% selecting 'cool'. This could be because of comprehension of the word 'cool', with students confusing it with 'cold', or 'to have a cold'. 'I feel cool' fits grammatically but does not fit with the context. Either way, it is nonetheless challenging to determine why the Early Childhood Education and Nutrition majors found this question more difficult. There are, naturally, other possible interpretations although the DIF analysis does not suggest why the DIF exists.

Alavi & Karami (2011) examined experts' interpretations of DIF results, arguing that while comparatively abundant research on the existence of DIF exists, little is dedicated to the interpretation of such results. They surveyed experts to determine if there was any agreement to the real cause of DIF and found that the experts opinions' were largely inconsistent. They conclude that DIF interpretations are ad hoc and that devising a mechanism for the interpretation of DIF results more systematically is required. Nonetheless, until a standardized process is developed, they warn that while DIF analyses should not be discounted results should always be carefully considered.

Conclusion

The goal herein however, is not to exhaust all possible explanations for significant DIFs. Rather, the ultimate purpose can be seen as a description of the process of a test evaluation. This evaluation produced useable results that may be able to provide direction for future improvement of the instrument, despite the fact that it is a low-stakes vocabulary test, designed in-house. This test was worth 15% of students' final grades, and

while no major decisions are based on the results, it could still mean a difference of a pass or fail in the course, and DIF analyses, especially when several academic disciplines are involved, should therefore not be overlooked.

References

Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc interpretations. TELL, 4(1), 1-18.

Alavi, S.M., Rezaee, A.A, & Amirian, S.M.R. (2011). Academic Discipline DIF in an English Language Proficiency Test. Journal of English Language Teaching and Learning, 5(7), 39-65.

Alderson, J. C., & Urquhart, A. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2, 192-204.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test, *Language Testing*, 27, 101-118.

Bond, T. G. & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: LEA.

Bond T.G., & Fox, C.M. (2007). (2nd ed.) *Applying the Rasch model: fundamental measurement in the human sciences*. Lawrence Erlbaum.

Bruckner, C., Saylor, M., Stone, W., Yoder, P. (2007). Construct validity of the MCD-1 receptive vocabulary scare can be improved: differential item functioning between toddlers with autism spectrum disorders and typically developing infants. *Journal of Speech, Language, and Hearing Research*, 50, 1631-1642.

Camilli, G. (2006) Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). New York: American Council on Education & Praeger series on higher education.

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, 59, 391-404.

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

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

Cole, N. S. (1993). *History and development of DIF*. Hillsdale, NJ, England: Lawrence Erlbaum Associates,Inc.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect

the mantel-haenszel and standardization measures of differential item functioning. In P. W. Holland, and H.Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and Modern Greek. *Language Learning*, 46, 233-282.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination. *Language Assessment Quarterly*, 4, 190-222.

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

Linacre, J. M. (2007). A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs. Chicago, IL: winsteps.com.

Linacre, J. M. (2004). Test validity and Rasch measurement: construct, content, etc. *Rasch Measurement Transactions*, 18(1), 970-971.

Linacre, J.M. (2008) *A User's Guide to Winsteps/Ministeps, Rasch Model Computer Programs*.

Linacre, J.M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA:Addison-Welsley Publishing Company.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, 22(4), 415-437.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Meade, A.W., & Fetzer, M. (2009).Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context. *Organizational Research Methods*, 12(4), 738-761.

Messick, S. (1989). Validity. In R.L. Linn (ed.) *Educational measurement* (pp. 13-103). New York: Macmillan.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing*, 21, 53-73.

Raju, van der Linde & Fleer, 1995). Raju, N. S., van der Linden, W. J. & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.

Rogers, H.J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning, *Applied Psychological Measurement*, 17(2), 105-116.

Runnels, J. (2011). Evaluation of an Achievement Vocabulary Test Using Rasch Analysis. *Studies in Linguistics and Language Teaching*, 22, 165-185.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2, 281-311.

Smith, E.V. (2000). Metric development and score reporting in Rasch measurement, *Journal of Applied Measurement*, 1, 303-326.

Smith, R. (2004). Detecting item bias with the Rasch model. *Journal of Applied Measurement*, 5(4), 430-449.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, 17, 323-340.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. London: The Guilford Press.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test:A theoretic comparison of methods. *Educational Measurement*, technical report No. 2.

Wolfe, E. W. (2004). Equating and item banking with the Rasch Model. In Smith, E. & Smith, R. (Eds.) *Introduction to Rasch Measurement* (pp. 360-390), Maple Grove, MN: JAM Press.

Wright, B.D., & Masters, G.N. (2002). Number of person or item strata, *Rasch Measurement Transactions*, 16, 888.

Wright, B. D., & Mok, M. M. C (2004). An overview of the family of Rasch measurement models. In Smith, E. & Smith, R. (Eds.) Introduction to Rasch Measurement (pp. 1-24), Maple Grove, MN: JAM Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

－平成24年11月 9 日　受理－