Evaluation of an achievement English vocabulary test using Rasch analysis

Evaluation of an Achievement English Vocabulary Test Using Rasch Analysis

Judith Runnels

Abstract

The Rasch model has recently been used in educational measurement as an evaluative tool for the validity of tests. Rasch analyses have been shown to map onto the six aspects of Messick's (1989) construct validity and compared to more classic models of test theory, make stronger arguments in providing validity evidence for tests. The Rasch model estimates the probability of a specific response according to person ability and item difficulty parameters, placing both on an interval scale. In the current study, an 83 item multiple-choice English vocabulary achievement test was administered to second-year non-English majors at a Japanese university. The test was developed from a 250 word study list. The results were analysed using a combination of Rasch measures and deterministic statistics, including logistic regression. The analyses highlighted several test items that exhibited unusual response patterning and suggested that the test was not an effective tool in measuring how well students acquired the 250 words on the study list. Deterministic and Rasch analyses were both effective as evaluative tools, although Rasch provided more precise information that can subsequently be used by test developers or educators to revisit potentially problematic test items, ultimately improving the validity of the test.

Key words: Rasch model, vocabulary test, validity, fit statistics, evaluation

Introduction

Increasing vocabulary is an essential part of language learning and taking

measurement of an individual's knowledge is a common goal of tests such as the Vocabulary Levels Test (Nation, 1983) and the Eurocentres Vocabulary Size Test (Meara & Buxton, 1987). Both of these tests are frequently used as a placement tool, as a measure of language learners' vocabulary size and sometimes as a diagnostic instrument to inform teachers where to begin their instruction (Nation, 1990). Evaluation of such tests has historically been done using deterministic measures (such as logistic regression) and more recently with Rasch-based approaches. Using these techniques for the analyses of pedagogical assessment has been shown to map onto the six facets of Messick's (1989) construct validity (content, substantive, structural, generalizability, external, and consequential) and used as a general criteria for providing validity evidence (Messick, 1995). For vocabulary tests specifically, this framework of evaluation has been used by Wolfe and Smith (2007), Beglar (2010), Beglar and Hunt (1999), Schmitt, Schmitt and Clapham (2001), Bond (2003) and Smith (2001), among others.

A primary goal of deterministic statistics is identifying significant differences across variables for sub-groups of test takers or items. One application of this, using logistic regression, can identify items that are biased. 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). When items are biased for a sub-group of test-takers, this is known as differential item functioning (DIF) and results in higher or lower scores for test takers within that group (Swaminathan & Rogers, 1990; 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). Logistic regression has been used as a check for DIF and DTF by Bruckner, Saylor, Stone and Yoder (2007), who detected differences in sub-groups of test-takers' responses to vocabulary multiple-choice questions (MCQ). It is important to consider this in the current analysis: since the test takers are all non-English majors, verification that the test is not biased towards one major or another is necessary.

In classic, deterministic models, a participant's overall raw score is assumed to be a measure of ability. A comparison of the responses to an individual item to overall scores on the test (known as the point bi-serial coefficient) is taken as a measure of item functioning (Cavanagh, Kent & Romanoski, 2005). This calculation can be across the test and sample population as a whole or among sub-groups of test takers or sub-groups of items. While this method is common, (see Bruckner *et al.*, 2007; Ackerman, 1992) it is also criticized: since vocabulary items are discrete, this type of analysis does not necessarily form a good representation of a test taker's vocabulary knowledge. In other words, the knowledge of one word does not necessarily predict the knowledge of another and it is therefore an inaccurate measure of ability (Schmitt, Schmitt & Clapham, 2001; Beglar, 2010). It does not necessarily follow to judge item validity according to responses to other items: using the point bi-serial coefficient as a measure of ability or as a judge of item functioning is limited (Schmitt, Schmitt & Clapham, 2001).

Rasch-based approaches, on the other hand use 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 consideration both item difficulty and person ability while assuming all test-takers to be independent (Rasch, 1980). Rasch analysis converts a test-taker'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). The same procedure is also applied to the probability that an item will be answered correctly. All logits are plotted on a single scale 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 and has the capability of identifying people or items that exhibit unexpected response patterns (Wright & Stone, 1979).

Most evaluations of vocabulary assessments have been performed on proficiency tests that aim to measure or estimate the size of an individual's vocabulary (see Beglar, 2010; Schmitt, Schmitt & Clapham, 2001). However, a proficiency test is not always necessary if the pedagogical goal is to determine how well the students acquired material presented in class. The current study on the other hand, is using the same types of analyses for an achievement test, with the aim of providing some preliminary evidence as part of an evaluation process of a newly developed MCQ vocabulary test. In order to do this, a study was designed using a combination of prescriptive (Rasch analyses) and deterministic statistics as an evaluative measure of a newly developed vocabulary achievement test. Deterministic statistics allow for comparisons across classes, question types or other groups and classifications as necessary (Schmitt, Schmitt & Clapham, 2001) while Rasch is able to provide detail about item difficulty, item functioning and test-takers' individual responses and ability. Combining the two methods and their associated analyses will provide useful evidence in the form of preliminary validity arguments in the evaluation of a test, whilst also allowing for updating any controversial items identified by the analyses.

Method

Participants

The test takers were 294 second year non-English majors, in 11 different classes organized according to their major (Early Childhood Education, Welfare Psychology, and Nutrition) from Hiroshima Bunkyo Women's University (aged 20 and 21 years old), a private university in Hiroshima City, Japan. All test-takers' first language (L1) is Japanese and second language (L2) is English.

Materials

The test consisted of 83 MCQs of 4 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). The questions were developed based on a 250 word study list that students received in the first week of the semester. All words and all sentences in the test items were taken directly from lesson materials, so it is assumed that the students had not only studied the list for the test but had also seen and/or used the words and sentences on the test during lessons. 91% of the words on the test fall within Nation's (2001) 3000 most frequent words of English and the remaining 9% were specific to the lesson content of the curriculum.

Procedures

The test was administered using www.classmarker.com©, an online testing site which automatically randomizes the correct response and distractors. Classmarker.com© does not allow any incomplete tests – a selection must be made

before the test taker may proceed to the following question. The test was taken in the students' usual class time and in their regular classrooms. Participants were allowed up to 90 minutes to complete the test. WINSTEPS Rasch software Version 3.72.4 (Linacre, 2008) and PASW Statistics 18, Release Version 18.0.0 (© SPSS, Inc., 2001, Chicago, IL, www.spss.com) were used to analyse the results of the test. Skewness, kurtosis and Cronbach's alpha for all scores were measured. Mean class scores, mean sub-group (for major) scores, mean scores across question type were compared using ANOVAs. As a check for DTF and DIF, ANOVAs were also performed for individual items, across all sub-groups (majors). Point bi-serial coefficients were calculated for all items on the test. Typically, items with a corrected-item total correlation (that particular item is not included in the calculation) value of near or less than zero should be explored and possibly removed or adjusted (an established acceptable range is 0.2 or higher; Churchill, 1979). A negative coefficient indicates that the response to the item contradicts the direction of the variable being measured and requires further evaluation (Churchill, 1979).

For the Rasch analyses, person-item maps, point-measure correlations, fit indices and unexpected responses were measured. The point-measure correlations, which refer to whether the responses to the item align with the abilities, should be positive so as to indicate a positive correlation with the average score of the other items (Wolfe & Smith, 2007). A negative correlation indicates that the item is functioning in direct opposition with the variable being measured and a near zero correlation means that the item was either very easy or very difficult to answer and may be confounding the results in some way (Linacre, 2007). The mean-square statistics (MNSQs) measure how well a test-taker's response patterning matches the predictions of the model (Smith, 2001). The

mean-square statistics (MNSQs) indicates the size of the misfit and the standardized z-score (ZSTD) indicates the significance level of the misfit (Bond & Fox, 2007). For a non-high stakes MCQ test, acceptable ranges for MNSQs are 0.7-1.3 and 0-2.0 for ZSTDs (Linacre, 2004).

The most unexpected responses were also measured. Unexpected responses are measured according to how the participant has performed overall, how difficult the individual item was for the population of test takers and are manifested in the form of standard residuals. The standard residual value illustrates the degree of unexpectedness of the response. Anything over a value of 5.0 is considered an extremely unexpected response implying either an issue with the test-taker or the item (Bohrnstedt & Knoke, 1982). These analyses are flagging questions that resulted in unusual response patterns.

Results

Descriptive statistics of the 294 tests were provided by PASW. The average test score was 86.2% (*SD* = 8.7%). The range of scores is shown in Figure 1 where it can be seen that scores are negatively skewed (skewness = 0.85, kurtosis = -0.58). Cronbach's alpha put the reliability of the test at 0.87 for the 83 questions.

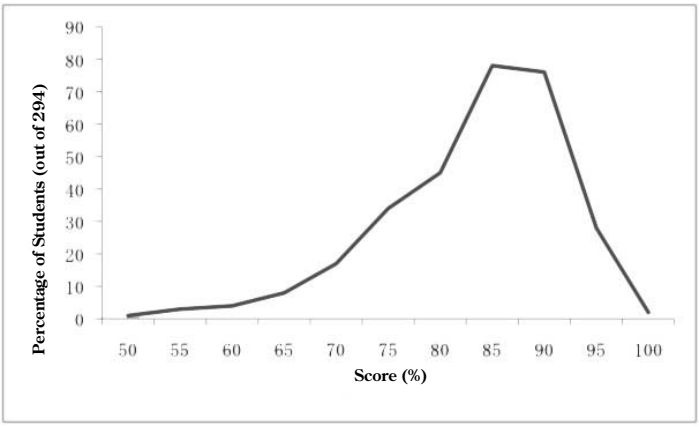

Figure 1. Distribution of student scores.

There were some significant differences across the 11 classes – the highest scoring class scored a mean of 94% (*SD*= 3.5%) and the lowest scoring class scored 78% (*SD* = 10.3%, F= 25.74, p = .00). There were no significant differences across sub-groups according to major – Early Childhood Education (*M* = 88.3%, *SD* = 7.3%), Welfare and Psychology (*M* = 82.9%, *SD* = 9.3%) and Nutrition (*M*= 84.3%, *SD* = 9.0%). This was the case for overall test scores and also for individual items (as a check for DTF and DIF). Furthermore, there were no significant differences for question type.

The Point Bi-Serial analysis identified 31 items with a correlation of less than 0.2 (Table 1). There were no negative correlations.

Table 1

Items with a Corrected Item-total Correlation (CIC) Under 0.2

Item	CIC	Item	CIC	Item	CIC	Item	CIC	Item	CIC
2	.000	19	.101	30	.149	49	.209	76	.173
4	.190	22	.114	32	.117	57	.178	77	.012
8	.191	23	.103	35	.000	65	.160	78	.147
10	.189	25	.000	37	.096	70	.151	81	.096
12	.109	28	.099	38	.174	71	.184	82	.076
16	.076	29	.000	46	.163	75	.092	83	.132
18	.021								

The results of the Rasch analysis are shown in Table 2. On the right of the y-axis are items, on the left of the y-axis are persons and to the far left are the Rasch linear measures in logits. When a person aligns with an item, this indicates that the person has a 50% chance of failure/success on that item. Several clusters of persons do not correspond to any item or difficulty level. Most of the items on the test have fallen below the clusters of people.

Table 2

Person-Item Map

```
                    Person - MAP - Item
                                 |
120                 #            +
                                 |
                                 |
                                 |
                                 |
                                 |
                ####             |
110                              +
                                 |
                                 |
                    T            |
                .###             |  Item 71
100                              +
                                 |
                .####            |
                                 |
                ######           |
             ############    S   |
90                               +T
                ########         |  Item 62
              ##########         |
                ########         |
                ######           |  Item 64
              ##########     M   |  Item 49
80              #####            +  Item 43
                .######          |  Item 46   Item 55
              .##########       |  Item 50   Item 65   Item 70
                .#####           |
              ##########         |  Item 44   Item 51   Item 52
70              #####            +S Item 33   Item 41   Item 59   Item 61
                .######     S    |  Item 63
                #####            |  Item 47
                  .##            |  Item 45   Item 57
                   .#            |  Item 11   Item 15   Item 54   Item 7
                  .###           |  Item 58   Item 74
60                .              +  Item 60   Item 68
                  #    T         |  Item 21   Item 72
                  .              |  Item 67
                  .              |  Item 13   Item 80
                  #              |  Item 53   Item 9
50                               +M Item 28
                  .              |  Item 26   Item 3    Item 42   Item 69
                                 |  Item 10
                                 |  Item 16   Item 18   Item 23   Item 48
                                 |  Item 14   Item 34   Item 40
                                 |  Item 82
40                               +  Item 37
                                 |  Item 39   Item 66   Item 79
                                 |  Item 1    Item 78
                                 |  Item 17   Item 31   Item 56   Item 6    Item 76
                                 |
                                 |  Item 22   Item 8
30                               +S Item 19   Item 24   Item 36   Item 5    Item 73   Item 81
                                 |
                                 |
                                 |  Item 20   Item 27   Item 75   Item 83
                                 |
20                               +
                                 |
                                 |  Item 12   Item 30   Item 32   Item 38   Item 4    Item 77
                                 |
                                 |
                                 |
10                               +T Item 2    Item 25   Item 29   Item 35
```

The results of the Rasch analysis for items are shown in Table 3. The items are arranged from most difficult to easiest. The first column, ' ENTRY NUMBER', corresponds to the test items (83 in total). 'TOTAL SCORE' indicates the total number of correct responses. 'TOTAL COUNT' is the total number of attempts and the 'MEASURE' column is the Rasch measure for this item (the difficulty; Wright & Panchapakesan, 1969) followed by the standard error. The infit and outfit statistics are in the next two columns, which show the MNSQ and the ZSTD. There are the point measure correlations and finally, the observed and expected scores.

The 'PT-MEASURE' and fit statistics columns are highlighted in vertical boxes in Table 3. There are no negative point-measure correlations. The 8 items with an observed point-measure correlations of a difference greater than 0.1 to what is predicted by the Rasch model are highlighted by horizontal boxes (Table 3). Any item that has a MNSQ or ZSTD outside of the acceptable range for infit is highlighted. For infit statistics, none of the items exhibit an MNSQ outside of the acceptable range, although 7 items (8.4% of the total items) fall outside of the acceptable ZSTD range. For outfit, twelve items (14.5% of total) fall outside of the acceptable range, 3 of which have significant ZSTDs (values over 2.0) – items 71, 46 and 49.

Table 3

Rasch Analysis for All Test Items

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
71	46	294	102.04	1.79	1.07	.6	1.56	2.3	.32	.40	84.6	86.1	Item 71
62	102	294	88.81	1.38	1.08	1.3	1.23	2.1	.38	.45	71.2	72.9	Item 62
64	131	294	83.55	1.32	.89	-2.1	1.00	.1	.52	.45	77.7	69.8	Item 64
49	148	294	80.60	1.31	1.17	3.2	1.37	4.1	.30	.45	63.7	69.2	Item 49
43	161	294	78.35	1.32	1.06	1.1	1.05	.6	.40	.45	67.8	69.6	Item 43
55	164	294	77.83	1.32	1.19	3.4	1.22	2.4	.30	.44	61.3	69.7	Item 55
46	166	294	77.48	1.32	1.25	4.4	1.32	3.4	.26	.44	59.6	69.8	Item 46
50	172	294	76.43	1.33	1.06	1.1	1.04	.5	.40	.44	67.8	70.1	Item 50
70	176	294	75.72	1.33	1.26	4.6	1.27	2.7	.25	.44	58.9	70.4	Item 70
65	177	294	75.55	1.33	1.22	3.9	1.27	2.7	.27	.44	60.6	70.4	Item 65
52	195	294	72.26	1.37	.93	-1.3	.83	-1.6	.49	.42	74.7	72.4	Item 52
51	198	294	71.69	1.38	.92	-1.3	.83	-1.6	.48	.42	74.3	72.8	Item 51
44	200	294	71.30	1.39	.90	-1.6	.87	-1.2	.49	.41	74.7	73.2	Item 44
33	202	294	70.91	1.40	.96	-.6	.90	-.9	.45	.41	73.3	73.5	Item 33
59	204	294	70.52	1.40	.96	-.6	1.00	.0	.43	.41	75.3	73.9	Item 59
41	206	294	70.13	1.41	1.00	.1	1.02	.2	.40	.41	74.7	74.2	Item 41
61	208	294	69.72	1.42	1.05	.8	1.02	.2	.37	.41	71.9	74.6	Item 61
63	213	294	68.70	1.44	1.01	.2	.94	-.4	.40	.40	75.7	75.7	Item 63
47	224	294	66.32	1.50	.91	-1.2	.80	-1.3	.46	.38	82.2	78.2	Item 47
57	232	294	64.45	1.56	1.17	2.0	1.14	.8	.25	.37	76.0	80.3	Item 57
45	233	294	64.20	1.57	1.06	.8	.92	-.4	.34	.37	77.7	80.6	Item 45
11	235	294	63.71	1.59	.98	-.1	1.22	1.2	.37	.36	81.2	81.2	Item 11
54	235	294	63.71	1.59	1.04	.5	1.11	.6	.33	.36	78.4	81.2	Item 54
15	236	294	63.45	1.60	.80	-2.4	.61	-2.4	.52	.36	85.6	81.4	Item 15
7	240	294	62.41	1.64	.84	-1.8	.69	-1.7	.48	.35	84.6	82.6	Item 7
58	241	294	62.14	1.65	.86	-1.5	.68	-1.8	.47	.35	84.2	82.8	Item 58
74	242	294	61.87	1.66	1.05	.6	1.41	1.9	.30	.35	82.9	83.1	Item 74
60	247	294	60.45	1.72	1.01	.1	.91	-.3	.34	.34	84.6	84.6	Item 60
68	250	294	59.54	1.76	.91	-.8	.66	-1.6	.42	.33	86.3	85.6	Item 68
21	254	294	58.26	1.82	.80	-1.8	.50	-2.4	.49	.32	87.3	86.8	Item 21
72	257	294	57.23	1.88	.86	-1.1	.64	-1.5	.43	.31	88.4	87.7	Item 72
67	261	294	55.75	1.97	1.02	.2	1.04	.3	.27	.30	88.7	89.0	Item 67
13	264	294	54.54	2.04	1.10	.7	.85	-.4	.25	.28	88.4	89.9	Item 13
80	265	294	54.12	2.07	.93	-.4	.91	-.2	.33	.28	91.1	90.2	Item 80
9	267	294	53.24	2.14	.94	-.3	.84	-.4	.31	.27	91.8	90.9	Item 9
53	267	294	53.24	2.14	.83	-1.1	.48	-1.9	.42	.27	91.1	90.9	Item 53
28	273	294	50.21	2.37	1.11	.7	1.56	1.5	.13	.25	92.5	92.9	Item 28
26	274	294	49.63	2.43	.93	-.3	.56	-1.3	.32	.24	92.8	93.2	Item 26
42	274	294	49.63	2.43	.96	-.2	.73	-.7	.29	.24	93.5	93.2	Item 42
69	275	294	49.03	2.48	.98	.0	.59	-1.2	.29	.24	93.2	93.5	Item 69
3	276	294	48.40	2.54	.98	.0	.87	-.2	.25	.23	94.2	93.8	Item 3
10	278	294	47.04	2.67	1.02	.2	1.78	1.7	.18	.22	94.5	94.5	Item 10

16	280	294	45.53	2.84	1.11	.5	1.27	.7	.11	.21	95.2	95.2	Item 16
18	280	294	45.53	2.84	1.14	.7	1.51	1.2	.07	.21	95.2	95.2	Item 18
23	280	294	45.53	2.84	1.10	.5	1.13	.4	.14	.21	95.2	95.2	Item 23
48	281	294	44.70	2.94	.97	.0	.74	-.5	.26	.20	95.5	95.6	Item 48
14	282	294	43.80	3.04	.96	-.1	.62	-.8	.24	.19	95.9	95.9	Item 14
34	283	294	42.84	3.17	.96	-.1	.51	-1.1	.26	.19	96.2	96.2	Item 34
40	283	294	42.84	3.17	.94	-.1	.68	-.6	.24	.19	96.2	96.2	Item 40
82	284	294	41.79	3.31	1.06	.3	1.58	1.2	.10	.18	96.6	96.6	Item 82
37	286	294	39.37	3.67	1.04	.2	1.41	.9	.11	.16	97.3	97.3	Item 37
39	287	294	37.94	3.91	.97	.0	.52	-.9	.20	.15	97.6	97.6	Item 39
66	287	294	37.94	3.91	1.01	.1	.46	-1.0	.19	.15	97.6	97.6	Item 66
79	287	294	37.94	3.91	.97	.0	.51	-.9	.21	.15	97.6	97.6	Item 79
1	288	294	36.30	4.20	.97	.0	.41	-1.1	.21	.14	97.9	97.9	Item 1
78	288	294	36.30	4.20	.99	.1	1.39	.8	.13	.14	97.9	97.9	Item 78
6	289	294	34.38	4.58	.93	.0	.62	-.5	.18	.13	98.3	98.3	Item 6
17	289	294	34.38	4.58	1.04	.2	1.05	.3	.09	.13	98.3	98.3	Item 17
31	289	294	34.38	4.58	.90	-.1	.90	.0	.21	.13	98.3	98.3	Item 31
56	289	294	34.38	4.58	.91	-.1	.29	-1.4	.24	.13	98.3	98.3	Item 56
76	289	294	34.38	4.58	.98	.1	2.18	1.6	.14	.13	98.3	98.3	Item 76
8	90	294	32.05	5.10	.95	.1	.62	-.4	.16	.12	98.6	98.6	Item 8
22	290	294	32.05	5.10	1.03	.2	.56	-.6	.12	.12	98.6	98.6	Item 22
5	291	294	29.08	5.87	.93	.1	.28	-1.3	.19	.10	99.0	99.0	Item 5
19	291	294	29.08	5.87	1.00	.2	1.85	1.2	.08	.10	99.0	99.0	Item 19
24	291	294	29.08	5.87	.91	.0	1.10	.4	.16	.10	99.0	99.0	Item 24
36	291	294	29.08	5.87	.95	.1	.23	-1.4	.19	.10	99.0	99.0	Item 36
73	291	294	29.08	5.87	.92	.0	.28	-1.3	.19	.10	99.0	99.0	Item 73
81	291	294	29.08	5.87	1.02	.2	.76	-.1	.10	.10	99.0	99.0	Item 81
20	292	294	24.92	7.15	.89	.1	.10	-2.0	.21	.08	99.3	99.3	Item 20
27	292	294	24.92	7.15	.96	.2	.23	-1.4	.16	.08	99.3	99.3	Item 27
75	292	294	24.92	7.15	1.00	.2	1.16	.5	.08	.08	99.3	99.3	Item 75
83	292	294	24.92	7.15	.95	.2	1.60	1.0	.10	.08	99.3	99.3	Item 83
4	293	294	17.89	10.06	.96	.3	.11	-1.9	.14	.06	99.7	99.7	Item 40
12	293	294	17.89	10.06	1.00	.3	.29	-1.2	.09	.06	99.7	99.7	Item 12
30	293	294	17.89	10.06	.98	.3	.18	-1.6	.12	.06	99.7	99.7	Item 30
32	293	294	17.89	10.06	1.00	.3	.26	-1.3	.10	.06	99.7	99.7	Item 32
38	293	294	17.89	10.06	.97	.3	.14	-1.8	.13	.06	99.7	99.7	Item 38
77	293	294	17.89	10.06	1.02	.3	1.04	.3	.03	.06	99.7	99.7	Item 77
2	294	294	5.77	18.31	MINIMUM MEASURE				.00	.00	100.0	100.0	Item 2
25	294	294	5.77	18.31	MINIMUM MEASURE				.00	.00	100.0	100.0	Item 25
29	294	294	5.77	18.31	MINIMUM MEASURE				.00	.00	100.0	100.0	Item 29
35	294	294	5.77	18.31	MINIMUM MEASURE				.00	.00	100.0	100.0	Item 35

Standard residuals greater than a value of 5.0 for any unexpected response are shown in Table 4. These occur for 14 items (17% of total).

Table 4 *Items with Standard Residuals (SR) over /5.0/.*

Item	*SR*	*Item*	*SR*	*Item*	*SR*	*Item*	*SR*	*Item*	*SR*
6	-8.16	17	-9.40	28	-10.70	54	-7.01	78	-15.32
	-7.64		-8.74		-6.74				-9.22
			-8.16						-6.94
			-7.64						
10	-16.13	18	-9.66	31	-14.86	56	-6.76	82	-11.64
	-7.89		-7.62						-9.19
			-6.91						-9.19
			-6.91						
11	-10.51	19	-21.99	37	-13.14	73	-7.43		
					-8.59				
					-7.32				
					-6.81				
16	-8.51	23	-8.51	48	-10.07	74	-11.52		
	-6.91		-6.91						

Summarizing all analyses, what has now been created is a list of items and responses to those items that require revisiting for either their low point bi-serial correlations, differences between predicted and observed PT-MEASURE correlations, misfitting fit statistics and/or a highly unexpected response to the item (Table 5).

Table 5 *Potentially Problematic Items*

Statistical Analyses	*Flagged Items*
Point Bi-serials	2 4 8 10 12 16 18 30 32 35 37 38 46 49 57 65 70 71 75
PT-MEASURE	16 18 20 28 46 57 65 70
Fit Statistics	10 18 19 28 46 49 71 76 78 82
St. Residuals	6 10 11 16 17 18 19 23 28 31 37 48 54 56 73 74 78 82

Discussion

The results showed significant differences across the 11 classes of students, although whether this is a reflection of the teacher or pre-existing differences among students is unknown. More importantly, however, is that there were no items that were functioning differently for different sub-groups of test takers (no significant differences across majors overall or for any individual item) and thus, neither DIF, DTF or in general, item bias are a likely possibility, arguing for the structural aspect of construct validity (Baghaei & Amrahi, 2011). Furthermore, students did not find any one question type easier to answer than others and while Rasch methods ranked individual items according to difficulty, it is useful to know that all question types showed even difficulty overall.

Rasch's person-item map (Table 2) showed that there was a mismatch between clusters of test takers and item difficulty – most items fell below the ability of test takers. This is known as an item-targeting problem and it represents a low precision of measurement (Wright, Mead & Ludlow, 1980). A perfect fit to the Rasch model is one where, when all items are lined up for difficulty from easiest to hardest, test takers fall in even increments across the span of the test items. Ideally, there are no gaps since this potentially indicates that some domain of the assessed variable is not being measured by the test (Baghaei, 2008). In the current case, this means that there were not enough questions of high enough difficulty to test the acquisition of the original 250 word study list and that this test was partly measuring another dimension, perhaps pre-existing knowledge for many of the words. Gaps in the person-item map can also implicate consequential validity since it can suggest that the results are not based on a population of widely-abled students (Baghaei & Amrahi, 2011) or that the items on the test are

not appropriately representative of the content being tested (Smith, 2001). Fit statistics can be used to check the relevance of test content since misfitting items are suggested to be measuring a different construct and threatening the generalizability aspect of validity (Baghaei, 2009). For the current test, while none of the infit MNSQs were outside of the acceptable range, only 3 items (3.6% of the total) exhibited significant outfitting MNSQ values (items 71, 49, 46). Misfitting infit is a greater threat to validity than outfit, since the infit reports a misfit in the region where the item is supposed to provide its most useful measurement – where the person's ability lies (Linacre, 2007). The responses to these three items, all significant outfit misfits, more likely reflect an issue with the item itself. Item 71 exhibited an extremely low correct response rate (Table 3, 15.6% of test takers responded correctly) and was far beyond the ability of most test takers (Table 2) suggesting that the misfit was due to the item difficulty, whereas items 46 and 49 fall within the clusters of most test takers ability (Table 2) suggesting that the misfit was more likely due to some test-takers' individual responses.

The items highlighted by all analyses (shown in Table 5) fall across all question types and may have been flagged for reasons that relate to either the item itself or the test-taker. It can be seen that the point bi-serial correlation analysis (Table 1) flagged almost as many items as the Rasch analyses (Table 3), although the information gleaned from Rasch is far more precise as to why the item is potentially problematic. For instance, some questions caused highly unexpected responses (as illustrated by the high standard residuals in Table 4). Unexpected responses need addressing to provide arguments towards the content aspect of construct validity by ensuring that the items have technical quality and are at an appropriate reading level with unambiguous phrasing (Messick, 1996). In the case of the

test-taker, unexpected responses could be due to random guessing, using a 'memory trick' to remember the answer, pre-existing knowledge, a clerical error and so on (Linacre, 2007). Since the online software used in the current test required a response, examinees were forced to guess if they did not know an answer (Schmitt, 2000). Linacre (2004) suggests that guessing is often the culprit behind unexpected responses, particularly if it is an unexpected success. The role of partial knowledge in vocabulary testing, as measured by guessing strategies and cautiousness, could be a very interesting direction for this group of testers since capturing partial knowledge, especially for low-level learners would greatly advance future versions of tests. Personal interviews with examinees could potentially provide insight into any guessing strategies (Schmitt, Schmitt & Clapham, 2001).

In terms of the potentially problematic items, there are several options for follow-up: complete elimination of the item, changing some or all of the distracters or changing the question stem. Making these adjustments would likely contribute to the validity arguments for future versions of the test but all require further investigations. Distracter analysis determines whether the test taker is meaningfully distracted (Baghei & Amrahi, 2011). It would also ensure test takers are engaged with the items by providing evidence for how well the distracters are causing responses that match "the intended cognitive processes around which the distracters were developed" (Wolfe & Smith, 2007, p. 209). For example, in the case of Item 46 of the current test, which was flagged by the point bi-serial, point measure coefficient and fit statistics analyses (Table 6), it was found that 60% selected the correct option. The distracter selection rates were 33%, 6% and 0%: two of the distracters were either not at all, or hardly engaging test-takers. It is likely

that higher scoring students selected the incorrect option that received 33% of all responses, which according to Baghaei and Amrahi (2011) is a threat to the substantive aspect of construct validity.

Factor analysis is also lacking from the current study (to determine the unidimensionality of the test) as are comparisons with scores from other forms of English assessment (oral assessment, overall course grade) for the sake of predictive validity and generalizability.

Conclusion

Despite the lack of both a normal distribution for mean scores on the test and a range of difficulty across items, the results of the statistical analyses provide an initial evaluation of the validity of items on an achievement test. While this test requires further exploration, the results here show the value of evaluations of achievement tests. Using deterministic and stochastic measures can provide useful tools for educators looking to identify problematic items or increase the validity of their assessment and these procedures should not only be restricted to placement or proficiency tests.

References

Ackerman, T. (1992). A didactive explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement,* 29, 67-91.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions,* 22: 1, 1145-1146.

Baghaei, P. (2009). A Rasch-informed standard setting procedure. Rasch Measurement Transactions, 23: 2, 1214.

Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research,* 2, 1052-1060.

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*. 16, 131-162.

Bohrnstedt, G.W. & Knoke, D. (1982). *Statistics for social data analysis.* Itasca, IL: F.E.Peacock.

Bond, T. G. (2003). Validity and assessment: a Rasch measurement perspective. Metodologia de las Ciencias del Comportamiento 5:2, 179-194.

Bond, T.G., & Fox, C.M. (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah NJ: Lawrence Erlbaum Associates.

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.

Cavanagh, R.F, Kent, D.B., & Romanoski, J.T. (2005). Illustrative example of the benefits of using a Rasch analysis in an experimental design investigation. Paper presented at the Conference of the Australian Association for Research in Education: Sydney.

Churchill, G.A., (1979). A paradigm for developing better measures of marketing constructs, *Journal of Marketing Research,* 16:1, 64–73.

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.

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

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

Linacre, John M. (2008), Winsteps Rasch Measurement Computer Program. Chicago: winsteps.com.

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

Meara P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142–154.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50, 741–749.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 241-256.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12–25.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

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

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.*

Chicago: University of Chicago Press.

Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55–88.

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

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

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

Wolfe, E. W. & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement,* 8:2, 204-234.

Wright, B. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational & Psychological Measurement,* 29:1, 23-48.

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

Wright,B.D., Mead,R.J., & Ludlow,L.H. (1980). KIDMAP: Person-by-item interaction mapping. Research Memorandum No. 29. MESA Psychometric Laboratory, Department of Education, University of Chicago.