Analyzing a Vocabulary Test with Rasch Analysis

Sayaka Karlin ※

Abstract

In this paper, the validity of a vocabulary test for first year university students was investigated. The vocabulary test includes 35 English sentences, and students are instructed to answer the meaning of each target word in Japanese. By using Rasch analysis, several aspects of the vocabulary test and the influence of loan words on the vocabulary test were studied. The result showed that loan words involved in the vocabulary test influenced the validity of the vocabulary test because some students were already familiar with the vocabulary meaning or they could assume the meaning of the vocabulary. It also showed the difficulty level of the vocabulary test for the students. This study examined a vocabulary test for Japanese students and how loan words influenced the validity of the test.

Key words: Loanwords, vocabulary, validity, Rasch analysis

Introduction

This paper focuses on assessing the validity of a vocabulary test for first year university students who majored in management. Vocabulary is the most basic linguistic proficiency required for language learning, and it is important for teachers to be familiar with how to make valid vocabulary tests and how to assess students' vocabulary knowledge appropriately. This paper also focuses on the effects of loan words (Japanese words that have been borrowed from English) on the validity of a vocabulary test. Loan words can have positive or negative influences on L2 learning, and it is important to be aware of the effects on a test's validity. It is not easy deciding which vocabulary should be included on a test, but using Rasch analysis to study the validity of the test and learn about the effects of selected vocabulary items can be of benefit for teachers when making tests.

Literature Review

Importance of vocabulary

It is a widely-acknowledged fact that vocabulary learning is important for L2 learning. The more vocabulary learners know, the more learners can understand the meaning of English and more effectively read, write, listen, and speak English. According to Nation (Nation, 2013), and it is believed that an educated adult English speakers know at least 20,000 words or more (Lightbown & Spada, 2013).According to Nation (2013), L2 learners need to know 3,000 to 4,000 word families in order to get 95% coverage of a text, and they need to know 6,000 to 9,000 word families in order to get 98% coverage of a text. Therefore, learners need to learn vocabulary to a certain threshold for L2 comprehension. However, vocabulary learning requires continuous learning and it can be tedious for some learners because of its featureless learning process. Giving vocabulary tests is one way to motivate students to continue learning the target vocabulary. When deciding vocabulary for a test, test makers need to consider factors such as the purpose of the test and the knowledge of the stu-

※ Komazawa University
Author contact: educ0199@komazawa-u.ac.jp

dents to be measured (Nation, 2013).

Validity

In language testing, test makers analyze what a test score means, and take action on the inference of the results (Fulcher & Davidson, 2007). Traditionally, validity in testing means the test is effectively measuring its intention, and three main types of validity include criterion-oriented validity, content validity, and construct validity(Hughes, 1989; Cronbach &Meehl, 1955).Criterion-oriented validity refers to how a particular test is related to a criterion. Content validity refers to how elements of a test are relevant and representative of the construct for measurement, and construct validity refers to how well a test measures the construct that it was designed to measure, and how well it reflects a particular construct. AERA et al (1985) described validity as the most important factor in test evaluation, and it involves appropriateness, meaningfulness, and usefulness for inferences based on test scores, and validity is how the accumulated evidence supports the inferences based on the score.

Loanwords

Loanwords are words whose meaning has been borrowed from another language with little or no changes (Matras, 2009). MacGregor has suggested that nearly 10% of Japanese is constituted of loanwords, and most of them are English words (MacGregor, 2002). According to Daulton (2008), about half of the first 3,000 words of English exist in some form or other in other languages, and many English loanwords in Japanese are borrowed from high-frequency vocabulary. Thus, learners' knowledge about loanwords can be an effective way for learners to improve English comprehension (Daulton, 2008). Daulton also found in his study that a relationship exists between learners' English proficiency levels and their ability to recognize connections to loan words, and stated that loan words can be used to assist English learning for learners (2008). Additionally, Olah (2007) stated that more focus is necessary on teaching loanwords in school because some loanwords are difficult for students to understand, and also teaching loanwords can be an effective strategy when teaching English words. Therefore, if loan words are learned correctly, they can lower learners' learning burden because they are already familiar with the meaning of the words and it requires less effort to learn the words (Nation, 2013). However, there are also negative aspects to loan words. Norman (2012) stated that even though loan words can help learners learn the target language, there is a risk that incorrect meaning of loan words could prevent appropriate communication with native speakers.

Additionally, the loan words in Japanese and their English equivalents need to have reasonably the same meaning in order to make use of those loan words, especially considering some Japanese loanwords have different meanings in English (Daulton, 2008). In short, it is beneficial and useful for learners if they learn the shared English meaning of Japanese loanwords correctly. However, when loanwords differ in meaning, communication can be obstructed. As the effectiveness of teaching and learning loanwords is considered, it is also advantageous to investigate the effects of loan words on the validity of a vocabulary test. In this study, the validity of Japanese loanwords on a vocabulary test is examined. Also, the validity of the vocabulary test is analyzed, in general. The three research questions to be investigated are as follows:

Research question 1: How effective was the vocabulary test measuring the vocabulary knowledge of students?
Research question 2: Are loan words more likely to have fit problems when conducting a Rasch analysis?
Research question 3:Did the vocabulary test demonstrate acceptable validity?

Participants

There were 27 participants involved with this study. Participants were first year university students in central Tokyo. Of the 27 participants, 12were female and 15were male. Students all majored in management, and were required to take a basic reading class during their first year. The reading class was 90 minutes and was held once a week. There were 15 weeks per semester. The university administered a placement test before the students' enrolment, and the 27 participants were placed in the low-intermediate level.

Instrument

Vocabulary data used for the analysis was collected from the vocabulary portion of the final exam in the fall semester. In every lesson during the semester, students took a vocabulary quiz which included 15 to 20 new target words. These words were chosen from the university's word book and the reading textbook. For the vocabulary quizzes, students were required to write the Japanese meaning of each target word. Each target word was underlined in an example sentence. For the final exam, the vocabulary section followed the same style as the vocabulary quizzes in the lessons. Thirty-five target words were included in the vocabulary section of the final exam, and these words were randomly chosen from both the university's word book and the textbook, provided that they were studied during the fall semester.

Procedures

Students' answers for the 35 vocabulary items were checked, and the results were converted into a text file for the Rasch analysis. For the Rasch analysis, the constructed text file was for a dichotomous test, with the correct answer represented as A and incorrect answers represented as B. For this study, partially correct answers were considered to be incorrect, and also represented as B.

Analyses

For the analyses, the collected test data was processed Winsteps software (Linacre, 2017) in order to conduct the Rasch analysis. With the Rasch analysis, the investigation focused on item polarity, person polarity, the variable (Wright) map, and dimensionality.

Results

Item Polarity (measure)

With regard to measure, Table 1 shows that the most difficult item on the test was *surprise*, and only nine students answered that item correctly. Also, for the word *violent*, only ten students answered that item correctly. The word *surprise* was used as a transitive verb in the sentence I *surprised* them, but eight students wrote the meaning as I *was surprised (by them)*.With regard to the word *violent*, six students wrote the meaning as *violence* instead of *violent*.

Table 1. Item Polarity (measure)

Total Score	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Item
9	2.57	0.47	1.05	0.3	1.24	0.7	1 surprise
10	2.35	0.46	1.16	0.8	1.07	0.3	35 violent
14	1.52	0.45	0.81	-1.0	0.79	-0.8	21 permission
15	1.32	0.45	0.66	-1.9	0.58	-1.8	22 responsibility
15	1.32	0.45	0.90	-0.5	0.84	-0.6	25 participate
16	1.12	0.46	0.78	-1.1	0.89	-0.3	18 require
16	1.12	0.46	0.96	-0.1	0.91	-0.2	26 represent
17	0.90	0.46	1.20	1.0	1.16	0.6	4 nervous
17	0.90	0.46	0.97	-0.1	0.83	-0.5	30 pleasure
19	0.46	0.48	1.15	0.7	1.26	0.8	2 honest
19	0.46	0.48	1.19	0.9	1.13	0.5	3 natural
19	0.46	0.48	0.82	-0.7	0.66	-0.9	15 admit
19	0.46	0.48	0.98	0.0	1.01	0.1	27 vote
19	0.46	0.48	0.84	-0.6	0.82	-0.4	33 common
20	0.22	0.50	1.63	2.3	2.80	2.9	7 director
20	0.22	0.50	1.35	1.4	1.83	1.7	8 serious
20	0.22	0.50	0.83	-0.6	0.75	-0.5	28 crime
20	0.22	0.50	0.73	-1.1	0.90	-0.1	34 scary
22	-0.33	0.55	1.23	0.8	0.82	-0.1	11 furniture
22	-0.33	0.55	0.83	-0.5	0.51	-0.8	14 purpose
22	-0.33	0.55	0.93	-0.1	0.60	-0.6	16 avoid
22	-0.33	0.55	0.96	0.0	0.88	0.0	20 trust
23	-0.65	0.59	1.23	0.7	1.67	1.0	5 action
23	-0.65	0.59	1.29	0.9	2.33	1.6	12 opinion
23	-0.65	0.59	1.05	0.3	0.74	-0.2	13 punishment
23	-0.65	0.59	0.86	-0.3	0.83	0.0	24 burn
24	-1.04	0.66	1.02	0.2	2.17	1.3	6 character
24	-1.04	0.66	1.27	0.7	1.28	0.6	9 perfectly
24	-1.04	0.66	0.90	-0.1	0.55	-0.3	31 affect
25	-1.55	0.77	0.81	-0.2	0.34	-0.4	10 boss
25	-1.55	0.77	0.97	0.1	1.40	0.7	17 control
25	-1.55	0.77	0.87	-0.1	0.44	-0.2	19 respect
25	-1.55	0.77	0.65	-0.5	0.23	-0.6	23 village
25	-1.55	0.77	0.65	-0.5	0.23	-0.6	29 focus
25	-1.55	0.77	0.81	-0.2	0.34	-0.4	32 contain

ItemPolarity (fit)

With regard to item fit, item reliability was 0.71. Generally, reliability over 0.80 is considered good (Linacare, 2017). As the item reliability was only 0.09 lower than the threshold for good item reliability, it was determined that item reliability was in the acceptable range. According to Bond and Fox (2007),a reasonable range for item fit MNSQ (infit and outfit) for multiple choice tests is within the range 0.7 to 1.3. As shown in Table 2,there were two misfitting items based on infit MNSQ and six misfitting items based on outfit MNSQ, exceeding the fit MNSQ of 1.3.The two misfitting items for infit MNSQ were *director* and *serious*. Additionally, the six misfitting items for outfit MNSQ were *director, opinion, character, serious, action*, and *control*

Table 2. Item Polarity (fit order)

Total Score	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Item
20	0.22	0.50	1.63	2.3	2.80	2.9	7director
23	-0.65	0.59	1.29	0.9	2.33	1.6	12 opinion
24	-1.04	0.66	1.02	0.2	2.17	1.3	6character
20	0.22	0.50	1.35	1.4	1.83	1.7	8 serious
23	-0.65	0.59	1.23	0.7	1.67	1.0	5 action
25	-1.55	0.77	0.97	0.1	1.40	0.7	17 control
24	-1.04	0.66	1.27	0.7	1.28	0.6	9 perfectly
19	0.46	0.48	1.15	0.7	1.26	0.8	2 honest
9	2.57	0.47	1.05	0.3	1.24	0.7	1 surprise
22	-0.33	0.55	1.23	0.8	0.82	-0.1	11 furniture
17	0.9	0.46	1.20	1.0	1.16	0.6	4 nervous
19	0.46	0.48	1.19	0.9	1.13	0.5	3 natural
10	2.35	0.46	1.16	0.8	1.07	0.3	35 violent
23	-0.65	0.59	1.05	0.3	0.74	-0.2	13 punishment
19	0.46	0.48	0.98	0.0	1.01	0.1	27 vote
17	0.9	0.46	0.97	-0.1	0.83	-0.5	30 pleasure
22	-0.33	0.55	0.96	0.0	0.88	0.0	20 trust
16	1.12	0.46	0.96	-0.1	0.91	-0.2	26 represent
22	-0.33	0.55	0.93	-0.1	0.60	-0.6	16 avoid
15	1.32	0.45	0.90	-0.5	0.84	-0.6	25 participate
24	-1.04	0.66	0.90	-0.1	0.55	-0.3	31 affect
20	0.22	0.50	0.73	-1.1	0.90	-0.1	34 scary
16	1.12	0.46	0.78	-1.1	0.89	-0.3	18 require
25	-1.55	0.77	0.87	-0.1	0.44	-0.2	19 respect
23	-0.65	0.59	0.86	-0.3	0.83	0.0	24 burn
19	0.46	0.48	0.84	-0.6	0.82	-0.4	33 common
22	-0.33	0.55	0.83	-0.5	0.51	-0.8	14 purpose
20	0.22	0.50	0.83	-0.6	0.75	-0.5	28 crime
19	0.46	0.48	0.82	-0.7	0.66	-0.9	15 admit
25	-1.55	0.77	0.81	-0.2	0.34	-0.4	10 boss
14	1.52	0.45	0.81	-1.0	0.79	-0.8	21 permission
25	-1.55	0.77	0.81	-0.2	0.34	-0.4	32 contain
15	1.32	0.45	0.66	-1.9	0.58	-1.8	22 responsibility
25	-1.55	0.77	0.65	-0.5	0.23	-0.6	23 village
25	-1.55	0.77	0.65	-0.5	0.23	-0.6	29 focus

Notes :*Item separation = 1.57; Item reliability = .71*

Person polarity (fit)

With regard to person fit, person reliability was 0.79, only 0.01 lower than the threshold of 0.80 generally considered to be reasonable. Thus the person reliability was considered acceptable. There were 26 students included in the study, as shown in Table 3, and one student (student 27) was not included because she answered all questions on the test correctly, which resulted in the maximum measure for infit MNSQ and outfit MNSQ. Regarding misfitting persons with MNSQ over 1.3, it is shown in Table 2 that there were two infit MNSQ misfitting persons items and six outfit MNSQ misfitting persons. The two infit MNSQ misfitting persons were student 3 and student 4. Additionally, the six outfit MNSQ misfitting students werestudent 1, student 2, student 3, student 4, student 5, and student 6. Student 3 was misfitting for both infit MNSQ and outfit MNSQ likely because she was late for the final exam, and only answered one side of the double-sided vocabulary test, forgetting to turn over the page and check the opposite side. Thus, this unusual condition may have caused her to misfit.

Table 3. Person Polarity (fit)

Total Score	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Student
29	1.93	0.49	1.21	0.8	1.93	1.6	Student 1
32	2.85	0.64	1.03	0.2	1.83	1.1	Student 2
12	-0.83	0.39	1.55	2.9	1.54	1.7	Student 3
21	0.49	0.39	1.33	1.9	1.52	2.0	Student 4
29	1.93	0.49	1.03	0.2	1.37	0.8	Student 5
33	3.34	0.76	1.22	0.5	1.35	0.7	Student 6
32	2.85	0.64	1.12	0.4	1.28	0.6	Student 7
18	0.05	0.38	1.11	0.8	1.15	0.7	Student 8
26	1.31	0.43	1.08	0.4	1.00	0.1	Student 9
30	2.19	0.52	1.07	0.3	0.80	-0.1	Student 10
22	0.64	0.39	1.05	0.4	1.06	0.3	Student 11
22	0.64	0.39	1.05	0.4	1.02	0.2	Student 12
32	2.85	0.64	0.98	0.1	0.69	-0.1	Student 13
13	-0.68	0.39	0.96	-0.2	0.84	-0.5	Student 14
32	2.85	0.64	0.96	0.1	0.61	-0.2	Student 15
33	3.34	0.76	0.95	0.1	0.78	0.2	Student 16
30	2.19	0.52	0.87	-0.3	0.93	0.1	Student 17
30	2.19	0.52	0.91	-0.2	0.60	-0.6	Student 18
21	0.49	0.39	0.90	-0.6	0.78	-0.9	Student 19
26	1.31	0.43	0.88	-0.5	0.65	-1.0	Student 20
27	1.50	0.45	0.87	-0.5	0.74	-0.5	Student 21
30	2.19	0.52	0.86	-0.3	0.86	0.0	Student 22
24	0.96	0.41	0.82	-0.9	0.78	-0.7	Student 23
30	2.19	0.52	0.81	-0.5	0.60	-0.6	Student 24
18	0.05	0.38	0.64	-2.6	0.58	-2.3	Student 25
19	0.20	0.38	0.63	-2.7	0.55	-2.4	Student 26

Notes :*Person real separation = 1.96; Person reliability = .79*

Variable (Wright) Map

The variable map displays higher-scoring to lower-scoring students along the left scale, as well as the difficulty of the words on the vocabulary test along the right scale. The variable map is seen Figure 1,with the ability level of students on the left side contrasted with the difficulty of items on the right side. One student, student 27, answered all questions correctly, which resulted in her being placed at the top of the left scale. Including student 27, there were seven students at the top of the left scale whose ability level was determined to be higher than the most difficult item. In addition, there were many items on the bottom of the right scale, falling below the ability level of the least-able student. The nine items that were below the ability level of all students in the study were *affect, character, perfectly, boss, control, respect, village, focus*, and *contain*.

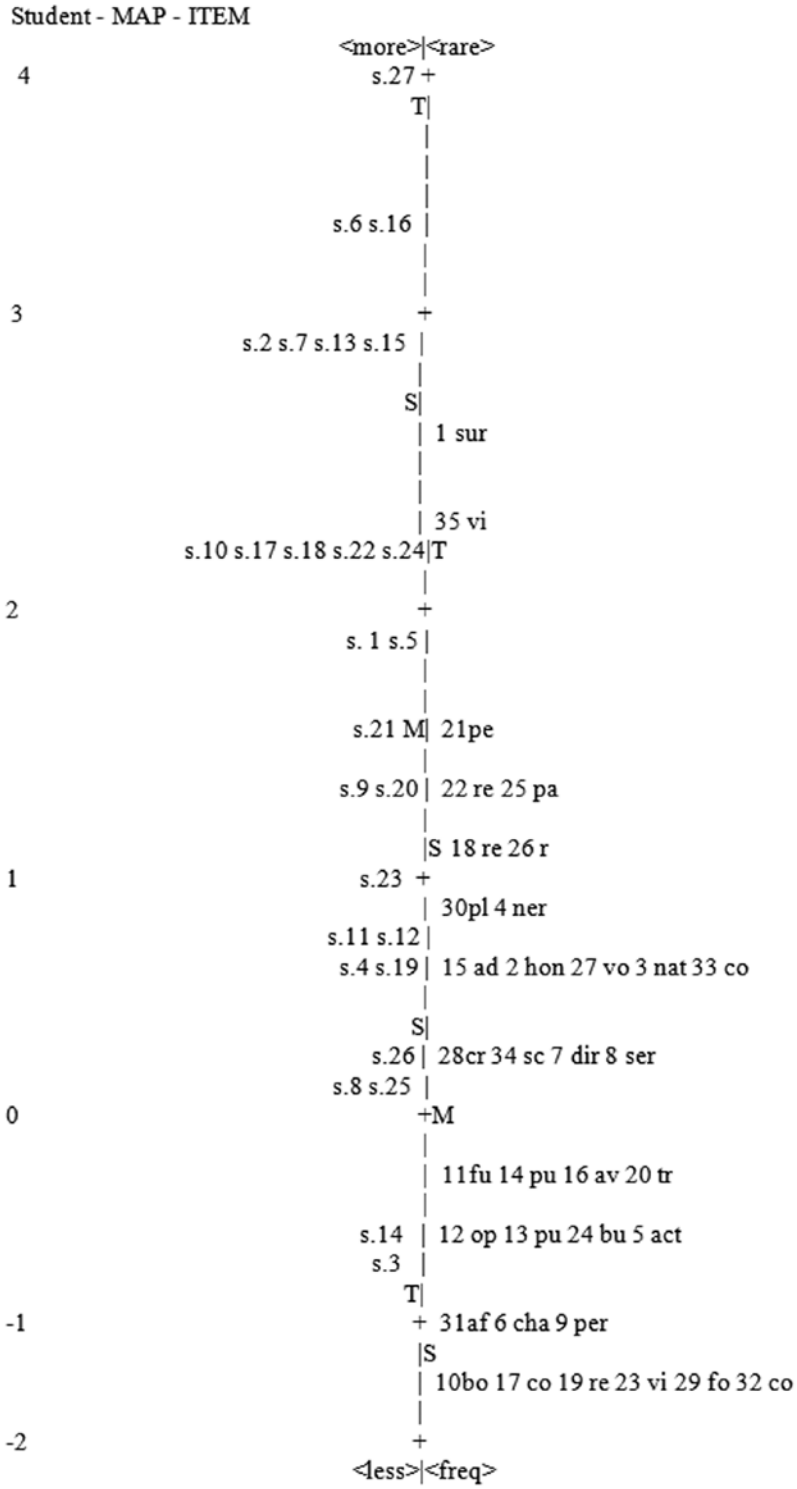

Figure 1. Variable (Wright) Map

Dimensionality Map

With regard to dimensionality, a test measuring a unidimensional construct should have residual unexplained variance in the first contrast of less than 3.0 (eigenvalue) and under 10% (unexplained variance). When these values are exceeded, it might be an indication of multidimensionality (Linacre, 2007). The standard residual variance is shown in Table 4, indicating possible multidimensionality with an eigenvalue of 4.80.

Table 4. Table of Standard Residual Variance in Eigenvalue Units

Eignevalue			Observed	Expected
Total raw variance in observations	=	51.1545	100.0%	100.0%
Raw variance explained by measures	=	16.1545	31.6%	31.0%
Raw variance explained by persons	=	8.3169	16.3%	16.0%
Raw Variance explained by items	=	7.8376	15.3%	15.1%
Raw unexplained variance (total)	=	35.0000	68.4%	100.0%
Unexplained variance in 1st contrast	=	4.8013	9.4%	13.7%
Unexplained variance in 2nd contrast	=	4.2156	8.2%	12.0%
Unexplained variance in 3rd contrast	=	3.4763	6.8%	9.9%
Unexplained variance in 4th contrast	=	2.9513	5.8%	8.4%
Unexplained variance in 5th contrast	=	2.5782	5.0%	7.4%

Figure 2 shows the standardized residual contrast 1 plot, and upper-case letters from A to R are placed abovethe center axis with lower-case letters p and q, while the remaining lower-case letters are placed below the center axis. Tables5 and 6 show the standardized residual loadings for all items, and represent the component analysis for two different contrasts. Table 5 includes 20 items, and Table 6includes 15 items.

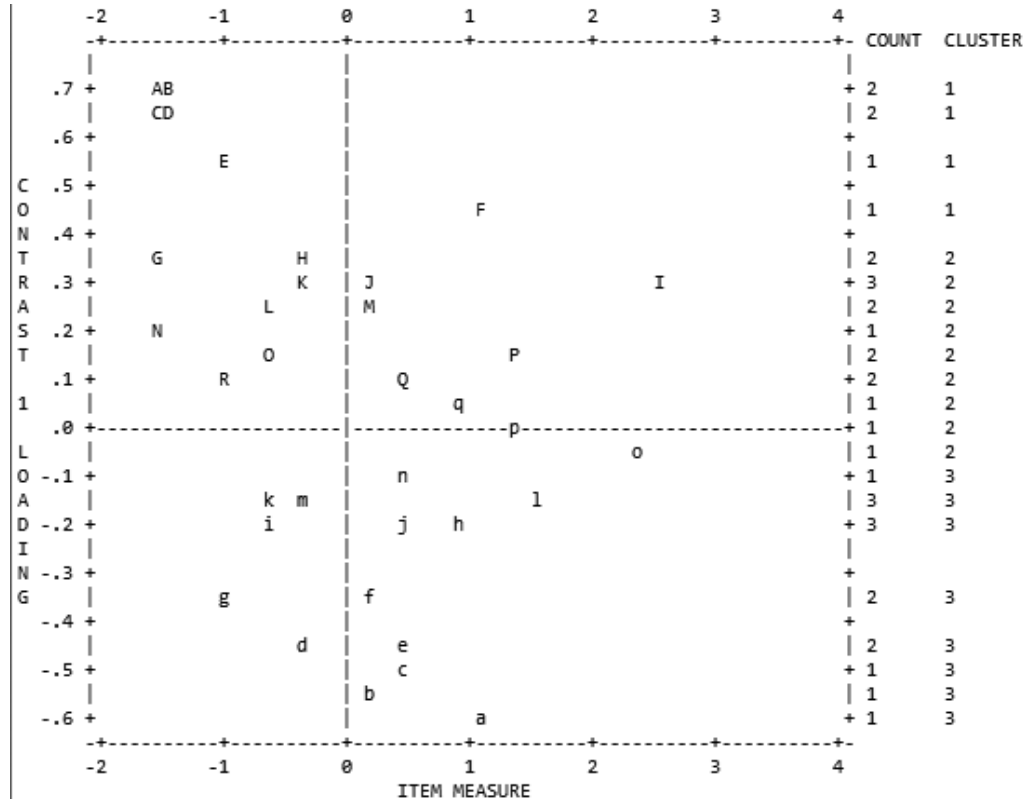

Figure 2. Standardized Residual Contrast 1 Plot

Table 5. Table of Standardized Residual Loadings for Items

Contrast	Loading	Measure	Infit MNSQ	Outfit MNSQ	Entry Number	Item
1	0.70	-1.55	0.81	0.34	A	10 boss
1	0.70	-1.55	0.81	0.34	B	32 contain
1	0.66	-1.55	0.65	0.23	C	23 village
1	0.66	-1.55	0.65	0.23	D	29 focus
1	0.53	-1.04	0.90	0.55	E	31 affect
1	0.43	1.12	0.96	0.91	F	26 represent
1	0.37	-1.55	0.97	1.40	G	17 control
1	0.33	-0.33	0.96	0.88	H	20 trust
1	0.31	2.57	1.05	1.24	I	1 surprise
1	0.31	0.22	0.83	0.75	J	28 crime
1	0.29	-0.33	0.93	0.60	K	16 avoid
1	0.25	-0.65	1.29	2.33	L	12 opinion
1	0.24	0.22	0.73	0.90	M	34 scary
1	0.19	-1.55	0.87	0.44	N	19 respect
1	0.16	-0.65	0.86	0.83	O	24 burn
1	0.14	1.32	0.66	0.58	P	22 responsibility
1	0.12	0.46	0.84	0.82	Q	33 common
1	0.08	-1.04	1.02	2.17	R	6 character
1	0.03	0.90	0.97	0.83	p	30 pleasure
1	0.02	1.32	0.90	0.84	q	25 participate

Table 6. Table of Standardized Residual Loadings for Items

Loading	Measure	Infit MNSQ	Outfit MNSQ	Entry Number	Item
-0.58	1.12	0.78	0.89	A	18 require
-0.55	0.22	1.35	1.83	B	8 serious
-0.50	0.46	1.19	1.13	C	3 natural
-0.47	-0.33	1.23	0.82	D	11 furniture
-0.44	0.46	1.15	1.26	E	2 honest
-0.35	0.22	1.63	2.80	F	7 director
-0.35	-1.04	1.27	1.28	G	9 perfectly
-0.20	0.90	1.20	1.16	H	4 nervous
-0.19	-0.65	1.05	0.74	I	13 punishment
-0.19	0.46	0.82	0.66	J	15 admit
-0.16	-0.65	1.23	1.67	K	5 action
-0.14	1.52	0.81	0.79	L	21 permission
-0.13	-0.33	0.83	0.51	M	14 purpose
-0.11	0.46	0.98	1.01	N	27 vote
-0.04	2.35	1.16	1.07	O	35 violent

Discussion

With regard to the first research question, *How effective was the test at measuring the vocabulary knowledge of students?*, Table 1 suggests that different types of verbs such as transitive or intransitive verbs, as well as word variants, may confuse students. Table 1 shows that only ten students correctly defined the meaning of violent, but six students wrote the meaning of *violence* instead. This may be because *violence* can sometimes be used as a loan word in Japan, so if students are familiar with the noun *violence*, they may have assumed that the meaning for the adjective *violent* was the same. Few students were able to differentiate between the

noun and adjective forms of the word.

With regard to the second research question, *Are loan words more likely to have fit problems when conducting a Rasch analysis?*, loan words did seem to have fit problems to a certain degree. In Table 2, there were six misfitting items, and four of these items were loan word items, specifically *director, character, action*, and *control*. These words are frequently used in Japan, and, as a result, some students of a lower ability could have unexpectedly correctly known the meaning of these words. If lower-level students acquire these loan words as they develop their L1, rather than through studying for an L2, the Rasch model likely would not be able to predict this result, causing fit to increase. Additionally, the variable map in Figure 1 showed that there were nine items, such as *affect, character, perfectly, boss, control, respect, village, focus*, and *contain*, which were considered to be easy for all students. Of these nine easy items, five were loan words. These loan words were *character, perfectly, boss, control*, and *focus*. It should be noted that the adverb *perfectly* is not used as a loan word in Japan, but the noun *perfect* is frequently used in Japan, so it can be assumed that students would have guessed the meaning of the word *perfectly* from the word *perfect* in this context.

The items for the two different contrasts are shown in Tables 5 and 6. Several factors may have separated these two contrasts, and one of them may have been the part of speech. In Table 5, there were 7 nouns, 11 verbs, and 2 adjectives. While in Table 6, there were 6 nouns, 3 verbs, 5 adjectives, and 1 adverb. More verbs were included in Table 5, and more adjectives were included in Table 6. With regard to loan words in Tables5 and 6, there were four loan words in Table 5;*boss*, focus, *control*, and *character*. In Table 6, there were also four loan words; *natural, director, perfectly, nervous and action*. In Table 5, there were two verbs and two nouns, but both *focus* and *control* can also be categorized as nouns, so this contrast could be interpreted as including four nouns. In Table 6, there were two nouns, two adjectives, and one adverb. Therefore, the different variants in these loan words might have influenced the separation of these two contrasts.

With regard to the third research question, *Did the vocabulary test demonstrate acceptable validity?*, the vocabulary test was mostly valid for the students. In Table 2, it was shown that the item reliability was 0.71, and in Table 3, it was shown that the person reliability was 0.79. As good reliability is considered to be 0.80, both of the reliability variables were within a reasonable range of being considered good. In Table 2, there were six misfitting items, but six misfitting items out of 35 is only 17% of the items. Also, in Table 3, there were six misfitting persons, but six misfitting persons out of 27 is only 22%) of the persons. Further, if only infit MNSQ is considered, which is the preferred measure of fit amongst researchers (Bond & Fox, 2007) there were only two misfitting items and two misfitting persons, indicating strong validity for this vocabulary test.

In Figure 1, it was shown that the proficiency level of students was widely variant. In this situation, it is not always easy to select vocabulary that meets all students' proficiency levels. Despite the variant proficiency level of the students on the variable map, it was also shown in Figure 1 that there were no vocabulary that matched the ability of seven high-level students on the test, and moreover, nine items on the vocabulary test were too easy for all students. Therefore, the variable map showed that the selection of vocabulary needs to be modified and provide better coverage of students' proficiency levels in order to improve the validity of the test. With regard to contrasts, it was shown that there were multiple contrasts on the vocabulary test. As is shown in Table 4, the eigenvalue for unexplained variance in the 1st contrast was 4.80, which exceeds the multidimensionality guideline of 3.0. However, Tables5 and 6showthat there was not significant multidimensionality found in the two different contrasts other than some word variants in the vocabulary, evidenced by suppressed residual loadings (only 11 items had a residual loading above +/- 0.40). According to Linacre (2007), if eigenvalue is more than 3.0, and variance is over 10%, multidimensionality might be present. In Table 3, eigenvalue for unexplained variance in the 1st contrast exceeded 3.0, but the variance was under 10% (9.4%). Considering this variance, multidimensionality was not strong on this test.

Conclusion

In this paper, the focus was on assessing the creation and validity of a vocabulary test for first year university students who majored in management in Tokyo, as well as the effects of loan words on the test. In the analysis, it was shown that several English words may confuse students because of variants associated with those words. Additionally, it was found that loan words influenced the validity of the vocabulary test. This may be because loan words were easier for students to learn compared with the target vocabulary. In addition, it can be assumed that students already knew the meaning of loan words without studying, or they assumed the meaning of loan words. Therefore, the fit and validity of the vocabulary items might improve if fewer loan words were included in the test. Additionally, the results showed that the vocabulary test was mostly appropriate for students' ability, but in order to improve the validity of the vocabulary test, several items need to be made more difficult so that the level of vocabulary items better matched with students' proficiency level. By using Rasch analysis, it is possible to see the validity of the vocabulary test as well as the adequacy of the vocabulary selection on the test. Moreover, the results from the analysis can be used in order to improve the validity of the test for the future use.

References

Bond, T.G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human science* (2nd ed.). Mahwah, NJ: Erlbaum.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loan words*. Clevedon: Multilingual Matters.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*.New York: Routledge

Hughes, A. (1989) *Testing for Language Teachers*,1st ed. Cambridge: Cambridge University Press.

Lightbown, P. M.,& Spada, N. (2013). *How languages are learned* (4th ed.). Oxford: Oxford university press.

Linacre J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago: MESA.

Linacre, J. M. (2017). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

MacGregor, L. (2003). The language of shop signs in Tokyo. *English Today*, 19(1), 18-23.

Matras, Y. (2009). Language Contact. Cambridge: Cambridge University Press.https://doi.org/10.1017/CBO9780511809873

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Ernst KlettSprachen.

Norman, J. (2012). Japanese university student awareness of waseieigo. *JALT 2011 Conference Proceedings*, 442-454.

Olah, B. (2007). English loanwords in Japanese: Effects, attitudes and usage as a means of improving spoken English ability. *Bunkyo Gakuin Daigaku Ningen Gakubu Kenkyū Kiyo*, 9(1), 177-188.

Winsteps.com. http://www.winsteps.com/winman/reliability.htm

Appendix A: Vocabulary section for the final exam

2017 Final Exam for fall semester

Section 1 - Vocabulary : Please write the Japanese meaning for the underlined words below (1.5 point / each).

No.	English words	Japanese meaning
1	I surprised them.	
2	She is a very honest person.	
3	Look at those mountains. The natural beauty of Nasu is magnificent.	
4	She was nervous about her presentation.	
5	His words and actions are not always the same.	
6	You cannot tell his character by his blood type.	
7	He is a famous movie director.	
8	Japan sometimes suffers serious damage from typhoons.	
9	She answered the question perfectly.	
10	I will ask the boss if I can have a day off.	
11	All our furniture is old.	
12	What is your opinion on this book?	
13	He received a very light and easy punishment.	
14	What is the purpose of your visit?	
15	The teacher admitted his mistakes.	
16	We must be careful to avoid traffic accidents.	
17	The driver could not control the car on the icy road.	
18	You are required by law to wear seatbelts in cars.	
19	I love and respect my parents.	
20	I have trust in him because he is sincere.	
21	Finally my parents gave me permission to go to America.	
22	We accepted responsibility for the mistake.	
23	I grew up in a small fishing village of 500 people.	
24	The toast is burning.	
25	You are invited to participate in the discussion.	
26	He is going to represent our school in the speech contest.	
27	Did you vote in the recent election?	
28	Japan has a low crime rate.	
29	The main focus of the talk is on the economy of Japan.	
30	She took great pleasure in entertaining them.	
31	The increase in CO_2 affects the global climate.	
32	This drink does not contain alcohol.	
33	Green tea is a very common drink in Japan.	
34	Scary things are happening one after another these days.	
35	The man was becoming violent, so she called the police.	