A Rasch-based Validation of The Word Associates Test

Tohru Matsuo

ラッシュモデルを用いた語彙連想テストにおける妥当性の検証

松　尾　　徹

Abstract

The purpose of this study is to validate Word Associates Test (WAT) designed to measure depth of lexical knowledge with Rasch Analysis. This study focuses on how the format which contains multiple correct items impacts the reliability and validity of the WAT. For this purpose, two versions of the test, varied (the number of correct items in each box varies) and the traditional four multiple-choice format were created. The results indicated that the traditional multiple-choice version yield better results than those of varied version in terms of good reliability, sufficient item fit, positive values of point measure correlations, and fundamentally unidimensionality of the measured construct. On the other hand, the varied version of WAT did not reliably distinguish between high and low ability of test takers' lexical knowledge. Furthermore, the varied version implied the high possibility of multidimensionality. That is, the WAT format is likely measuring two different constructs, polysemy and collocation.

Keywords: Word Associates Test, depth of lexical knowledge, Rash Analysis

（Received September 28, 2020）

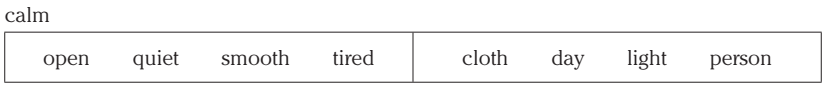
抄　　録

この研究の目的は語彙知識の深さを測定するために作成された語彙連想テストの妥当性をラッシュ分析を用いて検証することである。本研究では複数の正解がある語彙連想テストの形式が信頼性と妥当性にどのように影響するのかを調査する。この目的のために２種類の語彙連想テストが作成された。１つはそれぞれの枠の正解数が問題によって異なるもの、もう１つは各問題で正解数が１つしかない従来の４択問題である。分析の結果従来の選択問題形式の方が、信頼性、項目別の適応度の観点からより良いことが明らかとなった。一方で、正解が複数あるテスト形式は語彙の深さの知識が高い受験者と低い受験者を的確に分別できない場合が多いことが判明した。

キーワード：語彙連想テスト、語彙の深さ、ラッシュ分析

（2020 年 9 月 28 日受理）

Recent research has indicated the important role of lexical knowledge plays in both receptive and productive skills (Milton, 2009; Schmitt, 2010). Therefore, it is imperative to measure second language learners' lexical knowledge. Even though lexical knowledge is complex and multi-dimensional (Nation, 2001; Qian, 2002; Read 2004), one of the most prevailing distinctions is vocabulary size and depth (Schmitt, 2014). Vocabulary size refers to how many words learners know, whereas depth refers to how well learners know the individual words. While there are some tests designed to measure learners' vocabulary size such as Peabody Picture Vocabulary Test (Goulden et al., 1990) and Vocabulary Size Test (Nation 2008; Nation & Gu, 2007), few tests are utilized to measure depth of vocabulary knowledge. The test format that has been most utilized as a depth of vocabulary knowledge measure is the Word Associates Test (WAT). The Word Associates Test was developed by Read (1993, 1998, 2000) to measure two aspects of depth of vocabulary knowledge, paradigmatic knowledge and syntagmatic knowledge. The former is related to word meaning, especially polysemy and synonymy, and the latter is word collocation. The most recent version of the Word Associates Test (Read, 2000) contains a total of 40 items. Figure 1 shows a sample item.



Note. This figure illustrates the display of items on the Word Associates Test (Read, 2000).

Figure 1 ***Sample Item from the Word Associates Test***

Each item consists of one stimulus word, which is an adjective, and two four-word sets. One to three of the four words in the first set are synonymous with one aspect or the whole meaning of the stimulus word, whereas one to three words in the second set collocate with the stimulus word. Each item has a total of four correct answers. In the example item above, *calm* is the stimulus word. It is associated with *quiet* and *smooth* in the first set and collocates with day (i.e., *calm day*) and person (i.e., *calm person*) in the second set. Although the correct items are distributed evenly on both sides in this example, three possible situations exist in order to reduce successful guessing.

1. The two sets both contain two correct answers.
2. The first set contains one correct choice, while the second set contains three correct answers.
3. The first set contains three correct answers, while the second set contains only one correct answer.

Each correctly identified word is awarded 1 point; therefore, the maximum score is 160 (4 x 40 = 160). Even though two types of lexical knowledge, polysemy and collocation, are tested, this difference is disregarded in the scoring system. Because all the stimulus words are

adjectives, only knowledge of adjectives is tested; however, it is possible to argue that nouns are tested indirectly because test-takers must select the nouns that collocate with the stimulus adjective.

Gaps in the Literature

Even though the WAT has been widely employed by second language researchers (e.g., Greidanus et al., 2004; Qian 1999, 2002; Qian & Schedl, 2004) there have been relatively few studies directly validating the design and format of the WAT. Read (1998) conducted two pilot studies with students in an intensive EAP program in New Zealand to investigate the concurrent validity of the WAT. In the first study, 84 learners took the WAT and another test that required them to match words and definitions of the target adjectives utilized on the WAT. The Pearson correlation between the WAT and the concurrent measure was high at .86.

In the second study, 15 participants took the same two tests as in the first study and participated in an interview to elicit their knowledge of the target adjectives. The participants' responses were scored with the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996). Correlations indicated that the interview scores correlated more highly with a matching test (r = .92) than with the WAT (r = .76). Read (1998) argued that the format was affected by the test-takers' willingness to guess what the associates of the target word might be; thus, the WAT's susceptibility to guessing is one of the main threats to its validity.

Schmitt et al. (2011) conducted two validation studies on the WAT (1998) to investigate the design feature of the word associates format. They mainly employed a qualitative approach to answer part of the following questions:

1. What lexical knowledge does the WAT illustrate?
2. What is the best way to score the WAT?
3. What is the best way to interpret various WAT scores, especially split scores?
4. What strategies do examinees use when taking a WAT?
5. What effect does guessing behavior have on the WAT?

Eighteen Japanese adults studying English for academic purpose in western Japan participated in the first study. The participants consisted of undergraduate and graduate students along with several learners in a TOEFL study course. The participants had demonstrated mastery (90%) of the 2,000 level vocabulary on the Vocabulary Levels Test, and mastery or near-mastery of the 3,000 and Academic Levels. The researchers wrote a 40-item test of adjectives from the Academic Word List. After taking the WAT, the participants were interviewed to elicit an independent measure of their knowledge of 10 of the target words. The results of the interviews were compared with the WAT scores to determine how accurately the WAT represented

the test-takers' knowledge of the 10 target words. The results indicated that test-takers with scores of 0 to 1 had little knowledge of the word in the interview and those who scored 4 generally had good knowledge. However, split scores of 2 or 3 could indicate anything from no understanding to good knowledge of the target word. The researchers argued that while the WAT measures lexical knowledge fairly well at the extremes of the scale (0 or 4), interpreting split scores (2 or 3) is problematic (p. 109). Furthermore, the researchers found that in nearly half of the cases, the WAT appeared to overestimate the learners' knowledge of the target words, which provided support for Read's (1993) original concerns regarding the accuracy of item scores as measure of the target words.

In the second study, the participants were 28 international students studying at the University of Nottingham. The participants had studied English for at least eight years, and were considered to be relatively advanced L2 learners. Schmitt et al. (2011) employed the same approach as in the first study by examining WAT interview correspondences and the test-takers' strategy use. The results generally confirmed the findings of the first study. In addition, Schmitt et al. found that orthographically similar distractors such as *special* and *spur* (the target word was *spurious*) did not perform satisfactorily, whereas semantically related distractors such as *bland* and *boring* (the target word was *literal*) worked better with the six-option version and the unrelated distractors such as *active* and *fiction* (target word is *subordinate*) provided better results with the eight-option version. Furthermore, the researchers found that the 2-2 pattern produced the highest correlation between the word associates and interview results. On the contrary, they suggested that the 1-3 pattern was more susceptible to successful guessing because three nouns that can collocate with a particular target adjective tend to share a similar meaning.

Even though Schmitt, Ng, and Garras' study provided significant insights into the functioning of the WAT, they mainly employed a qualitative approach. While their study revealed learners' test-taking strategies and discrepancies between the test-takers' scores and lexical knowledge of polysemy and collocation of a limited number of words, they did not examine any statistical indices of item quality, which are crucial aspects of instrument validation. Furthermore, Schmitt et al. (2011) followed the conventional scoring system for the WAT without investigating the dimensionality of the test. That is, even though the WAT tested different aspects of word knowledge, polysemy and collocation, this difference was disregarded and each correctly identified item was awarded 1 point. This scoring system is valid only when polysemy and collocation are the fundamentally same construct. Therefore, it is imperative to examine the dimensionality of the WAT by providing validity evidence for the test through quantitative approaches that employ sophisticated statistical models such as the Rasch model (1960). To address these gaps, this study quantitatively examined the validity of WAT using the Rasch model. Specifically, the study focused on how the format containing

multiple correct items impacts the reliability and validity of the WAT. For this purpose, two versions of WAT were created. The first was a varied version, where the number of correct items in each side of the box is varied. This format is the same as Read's original WAT (1998). Another is a traditional four-multiple choice version, where there is only one correct answer.

Research Questions

1. How well does each item in the two versions of WAT fit the Rasch model?
2. How does the degree of unidimensionality differ between the two versions of WAT?
3. How well and precisely does each item in the two versions of WAT measure test-takers' polysemious and collocational knowledge?
4. What are the reliability and separation indices of each version of WAT?

Method

Participants

The current study consisted of two rounds of data collection at different times. The first data were collected from the 238 Japanese university law majors enrolled in one of the two classes *General English 1A* or *General English 2A*. All the classes were streamed based on the result of the Institutional Program Test of English for International Communication (IP TOEIC). Their English proficiency ranged from beginner to high-intermediate. In the second round of data, the participants were 104 university students, who were the same majors enrolled in the same classes. Their English proficiency was very similar to that of the participants in the first data collection.

The Instrument

The Word Associates Test (Read, 1998) was adapted in this study. First, for the purpose of examining word frequency effect, the five items (head words) were selected from each of the first 1,000 levels up to the sixth 1,000 word family level in British National Corpus. The word family is a more appropriate unit than the lemma because learners beyond minimal proficiency levels have some word-building knowledge and are able to understand that there is both a formal and a semantic relationship between regularly affixed members of a word family. Moreover, there is extensive research indicating that the word family is a psychologically real unit (Bertram et al., 2000; Bertram, Laine, & Virkkala, 2000; Nagy et al., 1989). Thus, the total number of the target items was reduced from 40 to 30 items (the total number of correct items was 4 x 30 = 120). Next, in order to investigate the impact of the format of the test, two versions of the WAT were created. The first version was a varied version, where the number of correct items in each side of the box is varied. The distribution patterns were the same as

those of Read's original WAT; three patterns (1-3, 2-2, 3-1) were utilized to distribute correct items. Another version was a traditional multiple-choice format, where there is only one correct answer among four choices. In this version, knowledge of polysemy and collocation was tested separately. Hence, in the knowledge of the polysemy test, test-takers were required to choose the option that shares the closest meaning to that of a headword. Figure 2 shows a sample item. The headword is *general* and the correct answer is *b, whole*.

general

a. closed
b. whole
c. different
d. same

Figure 2 ***Sample Item from The Traditional Multiple-choice Version of Word Associate Polysemy Test***

Likewise, in the knowledge of collocation test, test-takers were required to choose the option that collocates with the headword. Figure 3 shows a sample item. The headword is *general* and the correct option is *b, idea* (i.e., the collocation is *general idea*).

general_____

a. country
b. idea
c. return
d. street

Figure 3 ***Sample Item from The Traditional Multiple-choice Version of Word Associate Collocation Test***

Procedures

In the first part of data collection 1, the varied version of WAT was administered with 80 students. In the second part of data collection 1, the traditional multiple-choice format version of the polysemy test was administered to another 158 students who had not participated in the first part of the study, and in the following week, the same version of the collocation test was administered to the same participants who had taken the traditional version of the polysemy test.

In data collection 2, two versions of the WAT were randomly distributed to roughly the same number of the students ($N = 51$ for varied version and 53 for traditional version). The students, who took the traditional version, were instructed to complete the polysemy test first, and then the collocation test. They were neither allowed to look forward at the collocation section while answering items on the polysemy section nor look back at the polysemy test while answering items in the collocation test. Most students completed the test in about 30 minutes.

In sum, the total number of the participants for the varied version was 131 and the traditional version of polysemy and collocation test was 211, respectively. The data from the completed tests in both data collections were entered into an Excel spread sheet, exported to WINSTEPS 3.73 (Linacre, 2011), and analyzed using the Rasch dichotomous model (Rasch, 1960). The Rasch model provides a way to construct person measures and item difficulty estimates and which orders persons and items on the same interval logit scale (logarithm of odds unit). The Rasch model also provides a way to investigate the fit of items and persons to model expectations. For instance, misfitting items can indicate a bias or problems in the items, while test-takers who misfit the model can indicate numerous problems, such as not answering test items seriously. In addition, the Rasch model also provides a way to investigate the dimensionality of the data through analyzing the item residuals.

Definitions of Rasch Terms and Criteria for Rasch Validation Employed in This Study

Research Question 1 asked how well each item on the two version of WAT fit the Rasch model. To address this question, item fit, which indicates how well the items fit the Rasch model, was inspected. Two Rasch fit statistics are commonly utilized: infit and outfit mean-square (MNSQ) statistics. The item infit MNSQ statistic is sensitive to unexpected patterns by persons whose ability is at or near the item's difficulty estimate, whereas the item outfit MNSQ statistic is sensitive to the responses of persons far above or below the item's difficulty. Both infit and outfit mean-square indices range from 0 to infinity. Infit and outfit MNSQ criteria vary depending on *N*-size (Smith, Schumacker, & Bush, 1998); however, a value of 1.0 indicates that the data fit the Rasch model perfectly. A value greater than 1.0 (underfit) indicates that unmodeled noise or other sources of variance exist in the data, which degrade the precision of the measurement of the latent variable, whereas a value less than 1.0 (overfit) indicates that the model predicts the data too well, which causes lower error variances and inflated reliability estimates; however, overfitting items do not present the same threat to the precision of measurement as underfitting items.

Acceptable infit and outfit MNSQ ranges differ depending on the researcher. For instance, Bond and Fox (2007) suggested that MNSQ values between .70 and 1.30 are acceptable for dichotomous tests. Wright, Linacre, Gustafson, and Martin-Lof (1994) stated that acceptable values can fluctuate depending on the context of a test; on a high-stakes test, the appropriate range is 0.8-1.2, but for lower stakes tests, the acceptable range might be 0.4-1.2. The fit criteria used in this study is ± twice the standard deviation of the infit and outfit MNSQ statistics (McNamara, 1996) as the item and person fit criteria (McNamara, 1996) for two reasons. First, it is considered to be stricter compared to the aforesaid criteria, and second, it can customize the appropriate range for each data set. Therefore, different fit criteria are used for each version

of WAT. Regarding the procedure for inspecting item fit in this study, infit MNSQ over outfit MNSQ has been prioritized, as a small number of unexpected responses can have a big impact on outfit MNSQ, whereas they have relatively little impact on infit MNSQ (Bond & Fox, 2007). However, if an item slightly violated the infit MNSQ criterion, outfit MNSQ was examined to determine whether the item should be retained or not.

Research Question 2 asked how the degree of unidimensionality differed for each version of WAT. For this question, the dimensionality of the items hypothesized to measure the same construct was investigated through a Rasch Principal Component Analysis (PCA) of item residuals analysis. Several studies (e.g., Smith & Miao, 1994) have shown that the Rasch PCA of item residuals analysis is superior to traditional factor analytic approaches for assessing the dimensionality of instruments designed to produce a unidimensional measure of a latent variable. The Rasch model extracts the first major dimension in the data, which is the common variance among the items, and if the data are unidimensional and they fit the Rasch model, no systematic relationships should be present in the residuals. In this study, the following criteria from Linacre (2007) were used to investigate the dimensionality of items on the measured constructs:

- Variance explained by items > 4 x first contrast is good.
- Variance explained by measures > 50% is good.
- Unexplained variance explained by first contrast < 3.0 is good. Unexplained variance explained by first contrast < 1.5 is excellent.
- Unexplained variance explained by first contrast < 5% is excellent.

Research Question 3 asked how well and precisely each item in the two versions of WAT measured test-takers' polysemous and collocational knowledge. For this question, the Wright-map of each version of the WAT was examined to determine whether: (a) a sufficient number of items are included on the measurement instrument; (b) the empirical item hierarchy shows sufficient spread; and (c) gaps exist in the empirical item hierarchy.

Research Question 4 asked what the reliability and separation indices of each version of WAT are. To address this question, the Rasch item and person reliability and Rasch item and person separation estimates are reported. Rasch item reliability is an estimate of the replicability of item placement in a hierarchy of items along the measured variable, and Rasch person reliability is an estimate of the replicability of person placement that can be expected if the same respondents are given another set of items measuring the same construct. Person reliability is calculated as the ratio of adjusted true variance to observed variance and represents the proportion of variance that is not due to error.

Regarding the criteria for person and item reliability, the criteria provided by Fisher (2007)

was utilized in this study. According to Fisher, person and item reliability < .67 is poor, .67 to .80 is fair, .81 to .90 is good, .91 to .94 is very good, and > .94 is excellent. The item separation index is an estimate of the spread or separation of items on the measured variable whereas the person separation is an estimate of the spread or separation of persons on the measured variable. Compared with the Rasch person and item reliability estimates, these indices are more sensitive measures of reliability, as they are not bound by 1.00. A higher value indicates better separation. A desirable value for item separation is above 2.00, as this indicates that item difficulties cover a range of at least two statistically distinct groups.

Results

The Varied Version of WAT

The fit criteria were calculated using ± twice the standard deviations of the infit and outfit mean-square statistics (McNamara, 1996). As the standard deviation was .04, this resulted in a strict Infit MNSQ criterion of .92 – 1.08. Overfit, which indicates a possible violation of the assumption of local independence, was indicated by Infit MNSQ statistics under .92. The item 9 collocation (*change*) and 29 polysemy (*fierce*) slightly overfit the model with an Infit MNSQ statistic of .91. Underfit is viewed as a more serious problem than overfit because it indicates unexpected responses to an item by persons with ability estimates near the item's difficulty estimate. Underfit was indicated by Infit MNSQ statistics over 1.08. Four of the 120 items slightly underfit the model: item 12, collocation (*skin;* Infit MNSQ = 1.11), item 20, polysemy (*fixed;* Infit MNSQ = 1.11), item 22 polysemy (*stupid;* Infit MNSQ = 1.13). These four underfitting items were also problematic as their point-measure correlation index, which is the correlation between each item and the total score, were all negative. In addition to these four items, another seven items showed negative point-measure correlation; hence, a total of 14 items showed negative value of point-measure correlation. This implies that more able persons missed the item and the less able persons responded correctly. In other words, these items did not reliably distinguish between high and low ability test takers.

The dimensionality of the varied version of WAT items was examined through a Rasch principal component analysis (PCA) of item residuals using the criteria set by Linacre (2007). The variance explained by the items (12.0%) was not greater than four times the variance accounted by the first contrast (4.1%). Therefore, the first criterion was not met. The Rasch model accounted for 14.4% of the total variance (eigenvalue = 20.3), which was below the required the value of 50%. However, this was due to the relatively low person separation statistic of 1.08. The eigenvalue of the first residual contrast was 4.1, which was above the 3.0 criterion, so the third criterion was not met. The unexplained variance explained by first contrast (2.9%) was less than 5%, so the fifth criterion was met. Even though the fifth criterion

was met, the eigenvalue of the first residual contrast was above the acceptable criterion of 3.0. In addition, an inspection of the standardized residual contrast 1 plot was hard to confirm the fundamental unidimensionality of the construct; even though items are gathered diagonally the center but also concentrated horizontally on top end. Hence, further inspection is necessary.

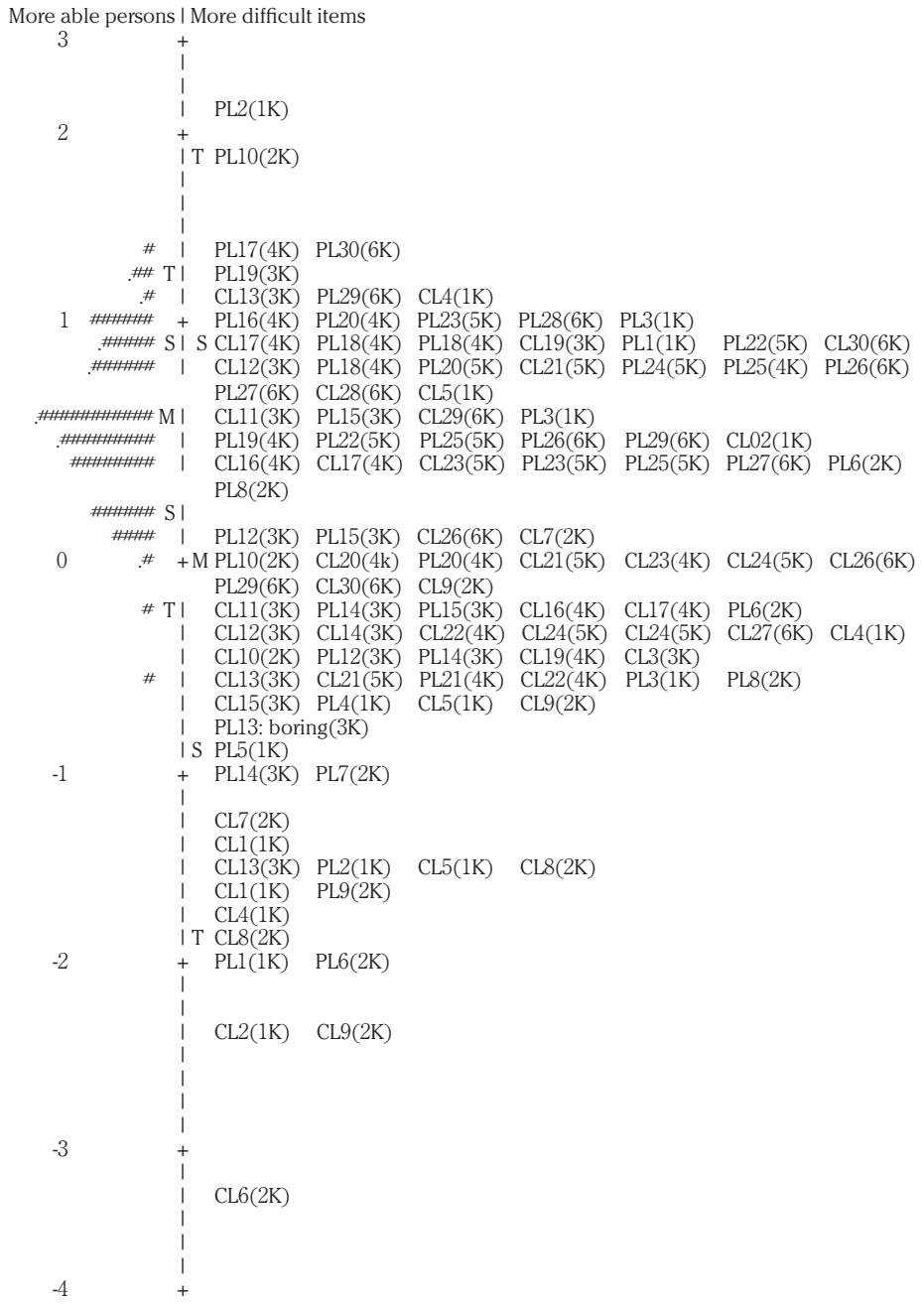
Figure 4 shows the linear relationship between the test-takers' abilities and the items' difficulties. On the far left side is the Rasch logit scale. Persons are indicated by '#' (representing two test-takers) and '.' (representing one test-taker). More able persons (i.e. higher-scoring persons) are toward the top of the figure and less able persons are toward the bottom. The Wright map for the varied version of WAT was examined to determine whether: (a) a sufficient number of items are included on the measurement instrument; (b) the empirical item hierarchy shows sufficient spread; and (c) gaps exist in the empirical item hierarchy. Figure 4 indicates that the varied version of WAT has a sufficient number of items, as the 120 items measure the full range of low and high-proficiency learners. No floor or ceiling effects were present for any examinees. The item mean is set to .00 (*SD* =.94) logits by convention, and the mean of the person ability estimates was .63 (*SD* = .36). Also, there were no significant gaps in the empirical item hierarchy, as items are found along nearly the entire measurement range.

The Rasch item reliability estimate was .94, which is excellent according to Fisher (2007), and the Rasch item separation index was 3.94. The Rasch person reliability was .54, which was poor according to Fisher (2007), and the separation index was 1.08, which was relatively low. This result was due to the participants' homogenous lexical proficiency. They had been screened by an entrance examination when they entered the university and they were subsequently streamed into their classes based on their TOEIC Bridge score.

The Traditional Multiple-choice Version of Word Associate Polysemy Test

The item fit was inspected using ± twice the standard deviations of the infit and outfit mean-square statistics (McNamara, 1996). As the standard deviation was .04, this resulted in a strict Infit MNSQ criterion of .92 – 1.08. Therefore, overfit was indicated by Infit MNSQ statistics under .92, and underfit was above 1.08. Neither overfit nor underfit items were identified. In addition, all point-measure correlation were positive; hence, all the items showed good fit to the model.

The dimensionality of the 4 multiple-choice version of WAT Polysemy items was examined through a Rasch principal component analysis of item residuals using the criteria set by Linacre (2007). The variance explained for by the items (19.7%) was greater than four times the variance accounted by the first contrast (4.4%). Therefore, the first criterion was met. The



Note. Each # equals 2 persons. Each . equals 1 person. M = Mean; S = One standard deviation from the mean; T = Two standard deviations from the mean. PL=Polysemy Test Item; CL= Collocational Test Item; (1K) = a head word from the first 1,000 word frequency level; (2K) = a head word from the second 1,000 word frequency level; (3K) = a head word from the third 1,000 word frequency level; (4K) = a head word from the fourth 1,000 word frequency level; (5K) = a head word from the fifth 1,000 word frequency level; (6K) = a head word from the sixth 1,000 word frequency level.

Figure 4 ***Wright-map for the 120 Items on the Varied Word Associates Test***

Rasch model accounted for 26.4% of the total variance (eigenvalue = 10.7), which was below the required value of 50%. However, this was due to the low person separation statistic of .86. The eigenvalue of the first residual contrast was 1.8, which was below the 3.0 criterion, so the third criterion was met. The unexplained variance explained by the first contrast (4.4%) was less than 5%, so the fifth criterion was met. Furthermore, an inspection of the standardized residual contrast 1 plot confirmed the fundamental unidimensionality of the construct. Hence, overall, the items appeared to form a unidimensional construct.

Figure 5 shows the linear relationship between 211 test-takers and 30 items. Figure 5 indicates that the 30 items of the traditional multiple-choice version of Word Associates Polysemy Test measure most of the range of low and high-proficiency learners. While no floor effects were present, slight ceiling effects were identified for two examinees. However, this was not problematic as there was no significant increase in the standard error of person ability estimates. The item mean is set to .00 (*SD* = 1.19) logits by convention, and the mean of the person ability estimates was -.14 (*SD* = .63).

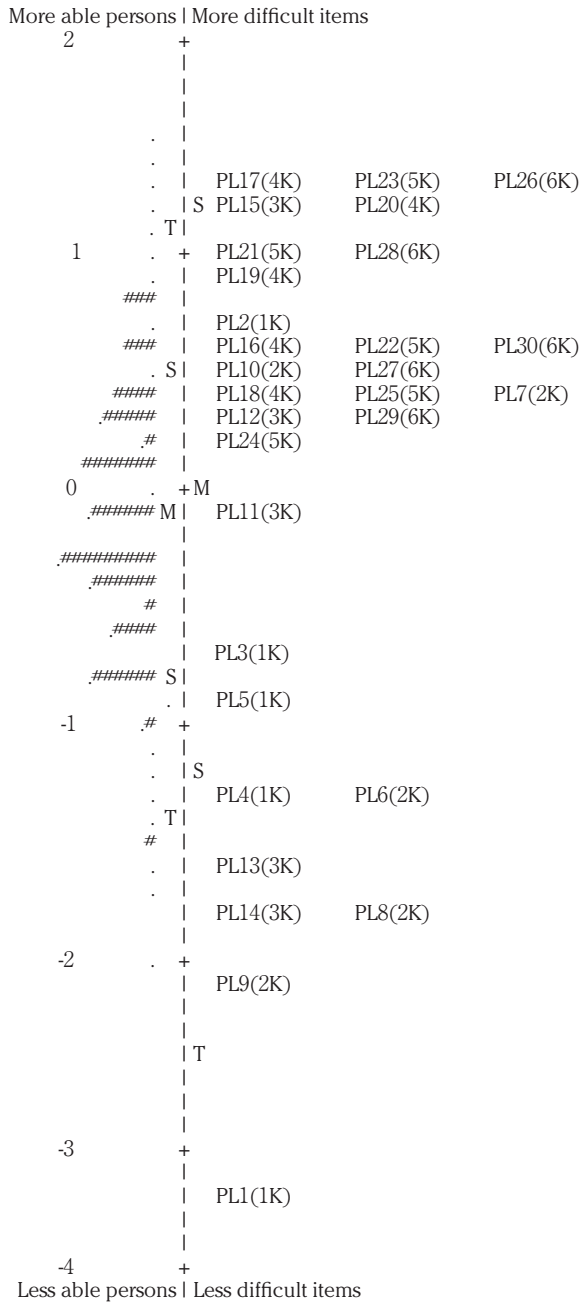
Figure 5 shows that no significant gaps exist in the empirical item hierarchy. Even though there is a gap between PL 11 (-.05) and PL 3 (-.65), there was no significant increase in the standard error of person ability estimates.

The Rasch item reliability estimate was .98, which is excellent according to Fisher (2007), and the Rasch item separation index was 6.72. The Rasch person reliability was .43, which was poor according to Fisher (2007), and the separation index was .86, which was low. This result was due to the participants' homogenous lexical proficiency. As it was described in the Rasch person reliability for the varied version of WAT, they had been screened by an entrance examination when they entered the university and they were subsequently streamed into their classes based on their TOEIC Bridge score.

The Traditional Multiple-choice Version of Word Associates Collocation Test

The item fit was examined using ± twice the standard deviations of the infit and outfit mean-square statistics (McNamara, 1996). As the standard deviation was .05, this resulted in a strict Infit MNSQ criterion of .90 – 1.10. Therefore, overfit was indicated by Infit MNSQ statistics under .90, and underfit was above 1.10. Item 15 (*curious;* Infit MNSQ = .89) was identified as overfitting. No underfitting items were identified. In addition, all point-measure correlations were positive; hence, all the items showed good fit to the model.

The dimensionality of the traditional multiple-choice version of WAT Collocation items was examined through a Rasch principal component analysis of item residuals using the criteria set by Linacre (2007). The variance explained by the items (18.3%) was not greater than four times the variance accounted by the first contrast (4.7%). Therefore, the first criterion was not met. The Rasch model accounted for 24.8% of the total variance (eigenvalue = 9.9),



Note. Each # equals 2 persons. Each . equals 1 person. M = Mean; S = One standard deviation from the mean; T = Two standard deviations from the mean. PL = the Traditional Word Associates Polysemy Test Item; (1K) = a word from the first 1,000 word frequency level; (2K) = a word from the second 1,000 word frequency level; (3K) = a word from the third 1,000 word frequency level; (4K) = a word from the fourth 1,000 word frequency level; (5K) = a word from the fifth 1,000 word frequency level; (6K) = a word from the sixth 1,000 word frequency level.

Figure 5 ***Wright-map for the 30 Items on the Traditional Multiple-choice Version of Word Associate Polysemy Test.***

which was below the required the value of 50%. However, this was due to the low person separation statistic of .81. The eigenvalue of the first residual contrast was 1.9, which was below the 3.0 criterion, so the third criterion was met. The unexplained variance explained by first contrast (4.7%) was less than 5%, so the fifth criterion was met. Furthermore, an inspection of the standardized residual contrast 1 plot confirmed the fundamental unidimensionality of the construct. Hence, overall, the items appeared to form a unidimensional construct.

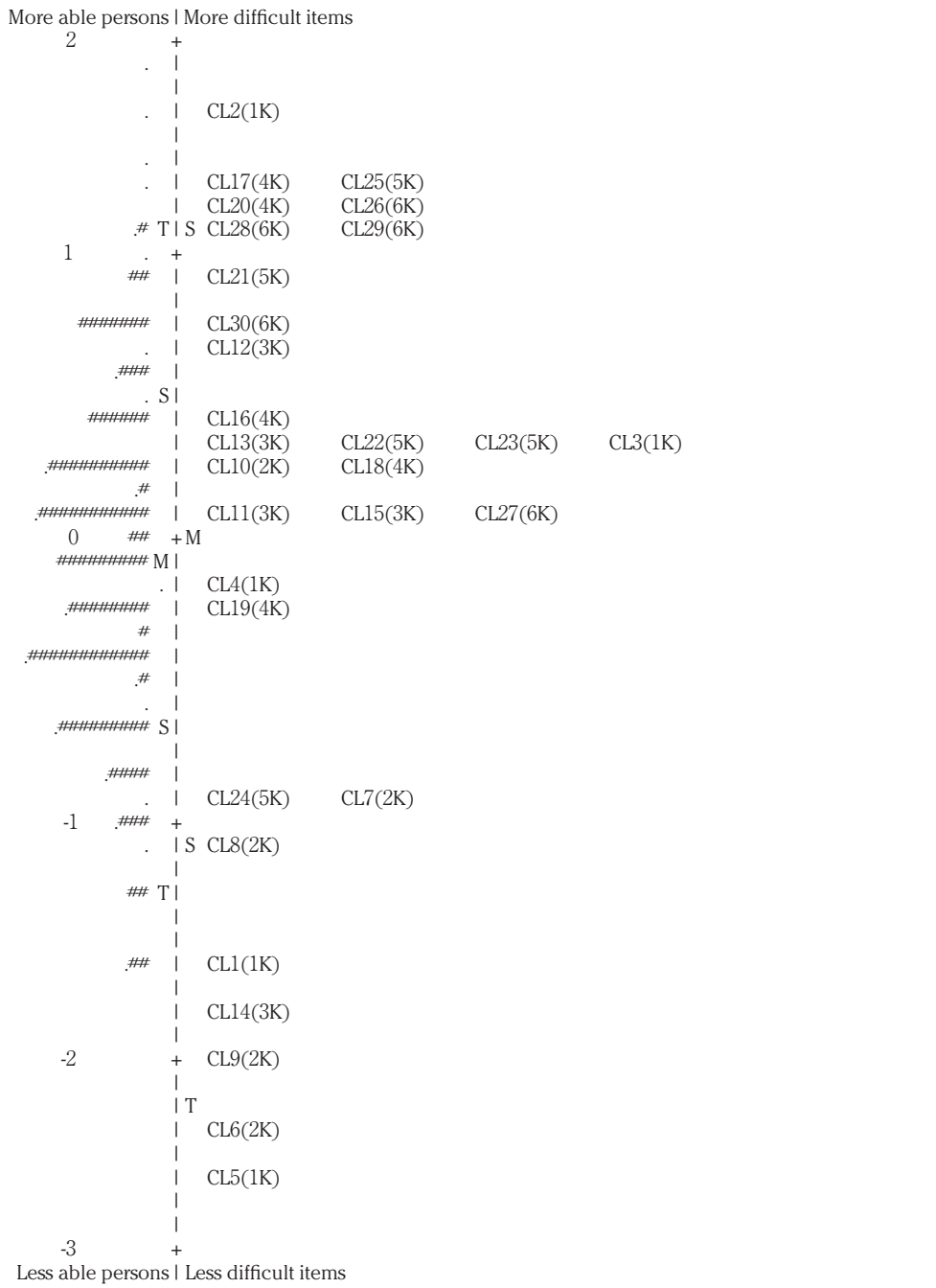
Figure 6 shows the linear relationship between 211 test-takers and 30 items. Figure 4 indicates that the 30 items of traditional multiple version of Word Associates Collocation Test measure most of the range of low and high-proficiency learners. While no floor effects were presented, slight ceiling effects were identified for one examinee. However, this was not problematic as there was no significant increase in the standard error of person ability estimates. The item mean is set to .00 (*SD* = 1.10) logits by convention, and the mean of the person ability estimates was -.09 (*SD* = .60).

Figure 6 shows that no significant gaps exist in the empirical item hierarchy. Even though there was a gap between CL 19 (-.24) and CL 7 (-.94), there was no significant increase in the standard error of person ability estimates.

The Rasch item reliability estimate was .98, which is excellent according to Fisher (2007), and the Rasch item separation index was 6.40. The Rasch person reliability was .39, which was poor according to Fisher (2007), and the separation index was .81, which was low. This result was due to the participants' homogenous lexical proficiency as they had been screened by the same examination and placement test.

Discussion

In this section, Rasch-based validation results of two versions of WAT were compared to determine which version yielded better validity and reliability. The first research question asked how well each item in the two versions of WAT fit the Rasch model. Table 1 shows the number of misfitting items and items with negative point-measure correlations in each version of WAT. As Table 1 indicates, the varied version of WAT has underfitting items and items with negative point-measure correlations, whereas the 4 multiple-choice version of both Word Association Polysemy and Collocation Test contain neither of these types of misfitting items. Among these misfitting items in the varied version of the WAT, the negative point-measure correlation is problematic. As it was mentioned in the Result section, this implies that more able persons missed the item and the less able persons responded correctly. Hence, these items on the varied version of WAT did not reliably distinguish between high and low ability test takers.



Note. Each # equals 2 persons. Each . equals 1 person. M = Mean; S = One standard deviation from the mean; T = Two standard deviations from the mean. CL = the Traditional Word Associates Collocation Test Item; (1K) = a word from the first 1,000 word frequency level; (2K) = a word from the second 1,000 word frequency level; (3K) = a word from the third 1,000 word frequency level; (4K) = a word from the fourth 1,000 word frequency level; (5K) = a word from the fifth 1,000 word frequency level; (6K) = a word from the sixth 1,000 word frequency level.

Figure 6 ***Wright-map for the 30 Items on the Traditional Word Associate Collocation Test***

Table 1 ***Summary of the number of misfitting items and items with negative point-measure correlation.***

Type of misfit	Varied	Polysemy	Collocation
Overfit Items	2	0	1
Underfit items	4	0	0
Negative PMC	14	0	0

Note. Varied = Varied version of WAT; Polysemy = traditional four multiple-choice version of Word Associates Polysemy Test; Collocation = traditional four multiple-choice version of Word Associates Collocation Test; PMC = point-measure correlations.

The second research question asked how the degree of unidimensionality differs between the two versions of WAT. Table 2 shows a summary of the Rasch PCA of item residuals of each version of the WAT. While the traditional multiple-choice version of WAT clearly indicated fundamental unidimensionality, the varied version of WAT is likely to form a multidimensional construct as the eigenvalue of the first residuals contrast of varied version was 4.1, which are above the acceptable criterion of 3.0.

Table 2 ***Summary of the Rasch PCA Item Residuals for Three Versions of WAT***

	Varied	Polysemy	Collocation
RV	20.3 (14.4%)	10.7 (26.4%)	9.9 (24.8%)
FC	4.1 (2.9%)	1.8 (4.4%)	1.9 (4.7%)

Note. Varied = Varied version of WAT; Polysemy = traditional multiple-choice version of Word Associates Polysemy Test; Collocation = traditional multiple-choice version of Word Associates Collocation Test; RV = Raw variance explained by the measure; FC = Unexplained variance in the first contrast.

In order to further investigate whether the varied version of WAT fundamentally measures a single construct, the polysemy and collocation items in the varied version of WAT were examined separately through a Rasch principal component analysis. Table 3 shows the results of the analysis for the varied version. As Table 3 shows, when polysemy and collocation items on the varied version were analyzed separately, the figure of first contrast decreased from 4.1 to 3.1, and 2.9, respectively. In addition, even though it did not change drastically, the Rasch model accounted for more of the variance for both polysemy and collocation. This result implied that polysemy and collocation are likely to be different constructs.

Table 3 ***Summary of the Rasch PCA Item Residuals for the Varied Version of WAT***

	Original	Polysemy	Collocation
RV	20.3 (14.4%)	11.4 (16.0%)	10.3 (14.6%)
FC	4.1 (2.9%)	3.1 (4.3%)	2.9 (4.1%)

Note. Original = original PCA analysis that polysemy and collocation were analyzed together; Polysemy = Polysemy items in varied version of WAT; Collocation = Collocation items in varied version of WAT; RV = Raw variance explained by the measure; FC = Unexplained variance in the first contrast.

The research question 4 asked what the reliability and separation indices of each version of WAT are. Table 4 shows the summary of the Rasch reliability and separation of two versions of WAT. Even though both versions of WAT indicated fairly good item reliability, the traditional multiple-choice version of Word Associates Polysemy and Collocation Test yielded better results. Furthermore, while item separation of the varied version was 3.94, that of the traditional version of Polysemy and Collocation Test was 6.72 and 6.40, respectively, which indicates that Traditional multiple-choice version of WAT has at least six levels of difficulty. It is worth noticing that the Varied version contained 120 items, whereas each Traditional multiple-choice test had only 30 items because knowledge of polysemy and collocation were tested separately and there was only one correct response among four choices.

Table 4 ***Summary of the Rasch Reliability and Separation of Two Versions of WAT***

	Varied	Polysemy	Collocation
Item reliability	.94	.98	.98
Item separation	3.94	6.72	6.40
Person reliability	.54	.43	.39
Person separation	1.08	.86	.81

Note. Varied = Varied version of WAT; Polysemy = traditional multiple-choice version of Word Associates Polysemy Test; Collocation = traditional multiple-choice version of Word Associates Collocation Test.

Conclusion

The purpose of this study is to validate two versions of WAT with Rasch analysis. Specifically, it investigates the impact of the format, focusing on how the different formats impact the reliability and validity of the WAT. It would be ideal that two versions of the WAT were randomly distributed to roughly equal number of students. However, this study was an initial validation study which employed Rasch analysis.

The results indicated that the traditional multiple-choice version worked better in terms of item reliability, positive point-measure correlations, and unidimensionality. Furthermore, the varied version of WAT showed negative point-measure correlations, which implied that the varied version of WAT did not reliably distinguish high and low ability test-taker's knowledge of polysemy and collocations. The results of a Rasch principal component analysis of item residuals indicated that polysemy and collocation, as measured by the varied version of WAT, were fundamentally different constructs. Therefore, they should be tested and scored separately.

References

Bertram, R., Baayen, R., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language,* 42(3), 390-405.

Bertram, R., Laine, M., & Virkkala, M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology,* 41(4), 287-296. doi:10.1111/1467- 9450.00201

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

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions,* 21(1), 1095.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics* 11(4), 341-363.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 191-208). Amsterdam, the Netherlands: Benjamins.

Linacre, J. M. (2007). *A user's guide to WINSTEPS.* Chicago, IL: Winsteps.com.

Linacre, J. M. (2011). WINSTEPS Rasch measurement computer program (version 3.73.0) [Computer software]. Chicago, IL: Winsteps.com.

McNamara, T. F. (1996). *Measuring second language performance.* London, England: Longman.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, England: Multilingual Matters

Nagy, W. E., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly,* 24, 263-282.

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

Nation, P., & Gu, P. Y. (2007). *Focus on vocabulary.* Sydney, Australia: National Centre for English Language Teaching and Research.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review,* 56(2), 282-308. doi:10.3138/cmlr.56.2.282

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning* 52(3), 513-536. doi:10.1111/1467-9922.00193

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing,* 21(1), 28-52. doi:10.1191/0265532204lt273oa

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedogogiske Instiut.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10(3), 355-371. doi:10.1017/S0142716400010602

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Erlbaum.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209-227). Amsterdam, the Netherlands: Benjamins.

Schmitt, N. (2010). *Researching vocabulary:* A vocabulary research manual. Basingstoke: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning,* 64(4), 913-951. doi:10.1111/lang.12077

Schmitt, N., Ng, J., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing,* 28, 105-126. doi:10.1177/0265532210373605

Smith, R. M. & Miao, C. (1994). Assessing dimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice,* Volume 2 (pp. 316-327). Norwood, NJ: Ablex.

Smith, R. M., Shumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement,* 2(1), 66-78.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review,* 53, 13-40.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* 8(3), 370. Retrieved from http://rasch.org/rmt/rmt83b.htm