A Rasch-based Validation of the Phrasal Vocabulary Size Test

Tohru Matsuo

ラッシュ測定モデルを用いた フレーズ語彙サイズ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 the Phrasal Vocabulary Size Test (PVST) designed to measure receptive knowledge of second language (L2) learners' formulaic language using Rasch Analysis. A total of 119 female university students in Japan participated in this study. The results indicated that most items in the PVST showed good fit to a Rasch model. The items covered the entire ability estimate range from the least able learner to the most able learners. In addition, the PVST showed both good item and person reliability estimates. Furthermore, the result of Rasch principal component analysis implied that the PVST measures a single construct. That is, even though the PVST consists of grammatically different types of formulaic sequences, they are fundamentally unidimensional of the measured construct.

Keywords: the Phrasal Vocabulary Size Test, formulaic language, Rash Analysis

（Received September 22, 2021）

抄　　録

この研究の目的は語句レベルの受容語彙知識を測定するために作成されたフレーズ語彙サイズテストがアジア圏の英語学習者、主に日本人学習者の語句レベルの語彙知識を測るテストとして妥当であるかについてラッシュ分析を用いて検証することである。分析の結果からこのテストのほとんどのアイテムがモデルに良く適合しており、語句レベルの語彙知識が低い学習者から高い学習者まで正確に測定できていることが明らかになった。また、ほぼ全てのアイテムが語句レベルの語彙知識が高い受験者と低い受験者を的確に分別できていることが判明した。さらに、このテストの目標アイテムは文法的に分類すると様々な種類の語句から構成されているが、ラッシュ分析で一次元性の度合いを調べた結果、これらは同じ構成要素（コンストラクト）である可能性が高いことが示唆された。

キーワード：フレーズ語彙サイズテスト、定型表現、ラッシュ測定モデル
（2021 年 9 月 22 日受理）

One of the most significant findings of corpus research shows that language consists of not only individual words, but also a great deal of units longer than a single word, which is commonly referred to formulaic language (Martinez & Schmitt, 2012). Formulaic language (FL) has been defined in numerous terms depending on researchers and research purposes. For example, Wray (2002) found over 50 terms to describe the phenomenon of formulaic language, such as formulaic sequence, chunks, multiword units, idiomatic expression, prefabricated routines, and many more. In attempt to create a consistent term in the field, Schmitt (2010) suggested FL as the umbrella term for the range of phrasal units that occur in language whereas formulaic sequence (FS) as the term for respective individual case of this phenomenon. This study follows this convention.

Recent research has shown the important role FL plays in language learning and use (e.g., Pawley & Syder, 1983; Wray, 2000, 2002). Pawley and Syder (1983) argued that FL enables native speakers of English to speak their first language fluently and to choose appropriate sequences of words that make them sound native-like. Native speakers have stored a large number of prefabricated chunks in their mental lexicon and when they wish to express a message, they can retrieve these chunks holistically instead of constructing them from individual words each time (Nation, 2013). Given the importance of FL in language learning and use, it is crucial to measure second language (L2) learners' knowledge of FL.

Despite a growing general interest in formulaic language, there have been no standardized test to measure L2 learners' knowledge of FL (Gyllstad & Shcmitt, 2019). Gyllstad and Schmitt (2019) argued that multiple categories of FL (e.g., idioms, collocation, lexical bundles, and phrasal verbs) and a very large numbers of formulaic sequences make it greatly challenging to develop a definite list of formulaic sequences and then develop a test based on the list. One of the few tests which is directly linked to the 505 formulaic sequences on the Phrasal Expressions List (Maritnez & Schmitt, 2012) was the Phrasal Vocabulary Size Test (Martinez, 2011b). This test contains the most common phrasal expressions made up of multiple categories in English, which could also potentially cause decoding problems for L2 learners if they read these phrasal expressions word by word.

Even though a prototype of the Phrasal Vocabulary Size Test (PVST) was thoroughly piloted with 2204 Austrian German speaking learners of English to select the 50 best items, the complete version of this test has not been further validated. Moreover, it has not been validated with L2 learners in other countries. Hence, the purpose of this study was to validate the PVST as a measure of Asian (mostly Japanese) learners' knowledge of

formulaic language using the Rasch model (Rasch, 1960).

Literature Review

In this section, firstly A Phrasal Expressions List (Martinez & Schmitt, 2012) is described as all items of the Phrasal Vocabulary Size Test that were selected from the list. Hereafter, the features of the PVST are described, and finally the previous validation study of this test is reviewed.

A Phrasal Expressions List

A Phrasal Expressions List (Maritnez & Schmitt, 2012) was created mainly for the purpose of being a) a guide for language learners and educators to employ formulaic language in their learning and teaching, specifically for receptive lexical knowledge, b) a means of including formulaic language in tests that assess receptive L2 knowledge and receptive skills, and c) an aid for monitoring the vocabulary acquisition process. The list consists of a total of 505 phrases that matched the frequency of words up to the 5,000 word frequency levels, which is considered a representation of the upper limit of general high-frequency vocabulary (Read, 2000). Martinez and Schmitt (2012) argued that this is a substantial number as the 505 multiword items would account for 10 per cent of the total items.

For compiling this Phrasal Expressions List, a two-step methodology was utilized to select phrasal words. The first step was an exhaustive computer-assisted search for co-occurring words with each frequency, statistical, and distributional data using the 100-million-word British National Corpus. The second step was to manually inspect those items with the guidance of pre-determined criteria, which consist of the following three aspects.

The first is that an expression should be a Morpheme Equivalent Unit (MEU). This means a phrasal item is processed as if it were a single morpheme, which is one definition of a phraseological lexical item (Wary, 2008). Martinez and Schmitt (2012) referred to the phrasal expression, *might* as *well* as an example of an MEU. The researchers argued that a learner who knows the meaning of this expression is unlikely to depend on form-meaning matching of the respective parts of the whole expression.

The second is that the expression should be semantically opaque. That is, the items included in the PHRASE list should be identified as causing difficulty for learners of English, specifically at the receptive level. For example, Martinez and Schmitt (2012) argued that the expression, *at this time,* might be categorized as an MEU since it means essentially the same as *now*. However, when a learner encounters this expression in the

text, even one who does not know the meaning of this expression can understand its meaning by simply adding up the meaning of individual words *at* + *this* + *time* (Martinez & Schmitt, 2012). Hence, this expression should not be included in the list.

The third is whether the expression is potentially deceptively transparent, which refers to words learners think they know but they do not (Laufer, 1989). Examples include *every so often,* which can be misunderstood as *often* (Martinez & Schmitt, 2012) and *at once*, which can be misunderstood as *first time*.

As the Phrasal Expression List contains the most frequent 505 phrasal expressions, it includes various types of formulaic sequences. One way to categorize these 505 expressions is based on grammatical types (Martinez, 2011a). As Table 1 shows, they can be grammatically categorized into six types: noun phrases (e.g., *point of view* and *well being*), verb phrases (e.g., *catch up* and *let alone*), adverbial phrases (e.g., *along with* and *on the way*), adjective phrases (e.g., *the odd* and *key to*), determiner/pronoun phrases (e.g., *the following* and *each other*), and other miscellaneous types such as interjection and other less frequent items (e.g., *oh dear* and *that is*). Of these six categories, Adverbial Phrase and Verb Phrase account for 84.15% of the entire phrasal expressions list items.

Table 1. ***Grammatical Analysis of Phrasal Expression List Items*** **(Martinez, 2011a, p. 159)**

Band	NP	VP	Adv	Adj	Det / Pro	Other
1K (k = 32)	0	7	23	0	1	1
2K (k = 85)	1	38	26	1	14	5
3K (k = 128)	2	45	63	3	12	3
4K (k = 158)	4	38	97	3	12	4
5K (k = 102)	4	32	56	2	5	3
Total = 505	11	160	265	9	44	16
Cum. %	2.17%	31.68%	52.47%	1.78%	8.71%	3.16%

Note. NP = noun phrase; VP = verb phrase; Adv = adverb; Adj = adjective; Det = determiner; Pro = pronoun; k = a number of items; 1K = the first 1,000 word frequency level; 2K = second 1,000 word frequency level; 3K = the third 1,000 word frequency level; 4K = the fourth 1,000 word frequency level; 5K = the fifth 1,000 word frequency level.

The Phrasal Vocabulary Size Test

The Phrasal Vocabulary Size Test (Martinez, 2011) was developed to measure L2 learners' receptive knowledge of common phrasal words and consisted of a total of 50 phrasal words items. The 50 phrasal words included in the Phrasal Vocabulary Size Test (PVST) are sampled from a total of 505 items in A Phrasal Expressions List (Martinez & Schmitt, 2012). The PVST was made up of 10 phrasal words per frequency level from the first to fifth 1,000 levels. The following is a sample item.

lead to: No one knows what it will **lead to**

a. want

b. have inside

c. cause in the future

d. find

As the sample item shows, all the target phrasal words are embedded in a simple non-defining context. As with the format of the Vocabulary Size Test (Nation & Beglar, 2007), the four-option multiple-choice format was utilized to (a) allow the test to be used with learners from a variety of language backgrounds, (b) control the level of item difficulty, (c) make marking as efficient and reliable as possible, and (d) make learners demonstrate knowledge of each item (Nation & Beglar, 2007). The test-takers are provided with the phrasal word form and have to access the meaning of the phrasal words. Test-takers have to have an adequately developed concept of the meaning of the target phrasal words to choose the correct answer from the four options because the correct answer and distractors usually share an element of meaning.

One of the advantages of the PVST is that the score can be clearly interpreted. As the items of this test were sampled from a finite selection of the phrasal expression list, the percentage correct on the test can be interpreted as the percentage known on the whole Phrasal Expression List, which is superior to other tests of formulaic language where there is no way to find out how to interpret the scores in terms of overall size (Gyllstad & Schmitt, 2019).

Previous Validation of the Phrasal Vocabulary Size Test

Martinez (2011a) validated a prototype of the Phrasal Vocabulary Size Test to select the best 50 items. The prototype of the PVST consisted of 15 target phrase items per word frequency band from the first 1,000 word to the fifth 1,000 word frequency level, for a total of 75 items. These items were allocated into three test versions. Version A contained items 1-6 of the respective word frequency levels, Version B 7-12, and Version C 1-3 (the same item as Version A) and items 13-15. Table 2 shows all the target items in each version of PVST. As Table 2 shows, the number of the target items in each word frequency level was six, for a total of 30 phrasal items in each test version.

A total of 2,204 participants, who were Austrian German speaking learners of English, took the test, with 742 taking version A, 731 Version B, and 730 Version C. All the test-takers were above 18 years old, and their English proficiency was around B2 Level in the Common European Framework of Reference for Language (CEFR). The means and standard deviation for the total scores (sum of all frequency bands) on all three versions

Table 2. ***Target items of each version of Phrasal Vocabulary Size Test***

Freq	Item	PVST A	PVST B	PVST C
1K	1	lead to	deal with	lead to
	2	have to	at all	have to
	3	a number of	be to	a number of
	4	go on	a lot	so that
	5	a bit	I mean	used to
	6	be likely to	at least	rather than
2K	1	as soon as	a range of	as soon as
	2	find out	as a result	find out
	3	so far	take place	so far
	4	to do with	and so on	in particular
	5	for instance	carry out	be expected to
	6	take over	each other	be about to
3K	1	it takes	feel like	it takes
	2	other than	or so	other than
	3	carry on	shake your hand	carry on
	4	all over	whether or not	give up
	5	turn out	get to	in touch
	6	in time	at once	get rid of
4K	1	as yet	in the light of	as yet
	2	prove to be	give rise to	prove to be
	3	in effect	no matter	in effect
	4	happen to	come across	might as well
	5	by no means	even so	next door
	6	take advantage	run out	on the one hand
5K	1	take for granted	keep on	take for granted
	2	as of	over time	as of
	3	would appear	come up to	would appear
	4	to blame	straight away	can tell
	5	stand for	shut up	under way
	6	by far	a handful of	turn down

Note. Freq = frequency; 1K = first 1,000 word frequency level; 2K = second 1,000 word frequency level; 3K = third 1,000 word frequency level; 4K = four 1,000 word frequency level; 5K = fifth 1,000 word frequency level; PVST = The Phrasal Vocabulary Size Test.

were calculated. Version A was 22.67 (*SD* = 5.30), Version B was 22.32 (*SD* = 5.76), and Version C was 19.95 (*SD* = 5.59).

Even though Version C seemed to be the most difficult of the three versions, the result of a one-way ANOVA showed that the difference among them was not significant. Reliability estimates were also examined for each version. Cronbach's Alpha for Version A was .869, Version B .854, and Version C .879, respectively.

For the criteria for selecting items, Martinez focused on inspecting the following four features, a) representativeness, b) difficulty, c) validity, and d) discrimination. Representativeness concerns whether the items of the Phrasal Vocabulary Size Test

represent the construct. Martinez (2011a) argued that all the target items reflected the criteria of semantic opacity. That is, a reading of each word would not indicate the meaning of the whole formulaic sequence. Therefore, the research concluded that respective target items could represent a measured construct.

Difficulty concerns whether the item is so easy as to be of limited value on the test. For examining item difficulty, the researcher used a *p*-value in classical item analysis, which was calculated using the total number of test takers who answer the item correctly divided by the total number of test-takers (Bachman, 2004). In addition to the calculation of difficulty of each target item, a *p*-value was utilized for distractor analysis (*p*-value was calculated with the same equation, but each distractor instead of the item as a dichotomous whole).

Validity concerns whether the item is measuring what it is intended to measure without evidence of extraneous, unintended linguistic, or non-linguistic influences. Discrimination concerns whether respective item can separate strong test-takers from weak ones. For examining the degree of discrimination, the researcher employed a score of point-biserial correlation, which was obtained by calculating a dichotomous item score (1 or 0) and its correlation with a total test score (Bachman, 2004).

The results of pilot tests revealed that the format accurately reflected true knowledge on the items tested. Moreover, the results indicated the good representativeness of the construct, and most items in the test were able to distinguish between test-takers with more phrasal vocabulary knowledge and ones with less.

Even though Martinez (2011a) made great efforts to validate The Phrasal Vocabulary Size Test in his pilot studies, his examination of statistical indices was limited in two ways. First, all the participants were Austrian German speaking learners of English. Therefore, generalizations about the response behavior in the items can be applied only to this particular group. As a result, the test should be validated with more learners of English in different countries.

Second, all the items of the test were analyzed through classical item analysis, which has three limitations. The first limitation of this analysis was that the results were only applied to this particular group. That is, generalizations regarding items and how they might behave with different groups are of limited validity. Second, the quality of item fit, a crucial aspect of instrumental validation cannot be investigated. Finally, the dimensionality of the Phrasal Vocabulary Size Test was not examined. Because The Phrasal Vocabulary Size Test is made up of different grammatical types of phrasal expressions, it is imperative to examine whether these different categories of phrasal expressions are fundamentally the same construct. To address these issues, this study examined the validity of The Phrasal Vocabulary Size Test using the Rasch analysis (Rasch, 1960), one

of the latent trait item response theory models.

Method

Research Questions

1. How well does each item in The Phrasal Vocabulary Size Test fit a Rasch model?
2. Does The Phrasal Vocabulary Size Test measure a single construct?
3. How well and precisely does each item in The Phrasal Vocabulary Size Test measure test-takers' phrasal vocabulary knowledge?
4. What are the reliability and separation indices of The Phrasal Vocabulary Size Test?

Participants

The participants were 119 female English-majors attending a private university in western Japan. There were 98 first-year students and 21 second-year students whose ages ranged from 18 to 20. The mean Institutional Placement TOEIC sores of the first-year university students were 465.00 (SD = 157.80) and the second-year students were 554.82 (SD = 147.22). Eleven students were international students, mainly from Vietnam and China, and the others (N = 108) were Japanese.

Instrument

The instrument was The Phrasal Vocabulary Size Test (BNC version 1-5K) (Martinez, 2011b), which was retrieved from Complete Lexical Tutor (Cobb, n.d.). The PVST was made up of 10 phrasal words per frequency level from the first to fifth 1,000 levels. Table 3 shows the target items of the test.

The platform of this test was changed from paper-based to an online version employing a Google form, which has a self-marking function.

Procedures

The 50-item Phrasal Vocabulary Size Test was administered to 119 Japanese university students during regular class time. All students took the test online using an iPad or smartphone. The lower proficiency students took 40 minutes and the intermediate proficiency students took 30 minutes to complete the test. The results were analyzed with the dichotomous Rasch model (Rasch, 1960) utilizing WINSTEPS version 3.73.0 (Linacre, 2011).

Table 3. ***Target Items of the Phrasal Vocabulary Size Test*** **(Martinez, 2011a)**

Freq	No.	Item	No.	Item
1K	1	go on	6	at least
	2	lead to	7	is likely to
	3	so that	8	is to
	4	at all	9	deal with
	5	I mean	10	used to
2K	1	so far	6	as a result
	2	to do with	7	as soon as
	3	take over	8	carry out
	4	in particular	9	be about to
	5	for instance	10	be expected to
3K	1	give up	6	all over
	2	feel like	7	in touch
	3	turn out	8	get rid of
	4	other than	9	at once
	5	get to	10	in time
4K	1	prove to be	6	in light of
	2	next door	7	by no means
	3	run out	8	come across
	4	take advantage	9	happen to
	5	in effect	10	even so
5K	1	by far	6	to blame
	2	come up to	7	take for granted
	3	straight away	8	as of
	4	would appear	9	can tell
	5	turn down	10	over time

Note. Freq = frequency; 1K = the first 1,000 word frequency level; 2K = the second 1,000 word frequency level; 3K = the third 1,000 word frequency level; 4K = the fourth 1,000 word frequency level; 5K = the fifth 1,000 word frequency level.

Results & Discussion

Research question 1 asked how well each item in the Phrasal Vocabulary Size Test fit the Rasch model. To address this question, item fit, which indicates how well the items fit the Rasch model, is inspected. Two Rasch fit statistics, infit and outfit mean-square (MNSQ) statistics, are commonly utilized. 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. Infit and outfit MNSQ criteria vary depending on *N*-size (Smith et al, 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.

The fit criteria were calculated using ± twice the standard deviations of the infit and outfit mean-square statistics (McNamara, 1996). The standard deviation for infit MNSQ and outfit MNSQ were .13 and .23, respectively; thus the fit criterion for infit MNSQ was .74-1.26 and that for outfit MNSQ was .54. -1.46. Table 4 displays a summary of the Rasch descriptive statistics for the 50 Phrasal Vocabulary Size Test items. The infit MNSQ statistics ranged from .78 to 1.38, the outfit MNSQ statistics ranged from .57 to 1.87, and the point-measure correlations were between -.08 and .62. The standard error ranged

Table 4. ***Rasch Descriptive Statistics for The 50 Phrasal Vocabulary Size Test Items***

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
PV29	2.03	.28	.87	-.7	.57	-1.6	.48
PV20	1.87	.27	1.21	1.2	1.87	2.7	.00
PV50	1.48	.24	1.20	1.4	1.34	1.5	.12
PV49	1.36	.24	.94	-.4	.85	-.8	.43
PV36	1.25	.23	1.01	.1	1.02	.2	.35
PV46	1.25	.23	1.30	2.3	1.40	2.0	.04
PV25	1.20	.23	1.21	1.7	1.18	1.0	.17
PV05	.90	.22	.89	-1.0	.86	-1.0	.49
PV22	.85	.22	.99	-.1	.96	-.2	.39
PV38	.81	.21	1.02	.2	1.00	.1	.37
PV13	.76	.21	.95	-.4	.98	-.1	.42
PV30	.76	.21	1.03	.3	.99	.0	.36
PV45	.67	.21	.86	-1.6	.90	-.8	.51
PV12	.63	.21	.98	-.2	1.07	.6	.38
PV47	.54	.21	.99	-.1	1.04	.4	.38
PV48	.54	.21	1.07	.8	1.12	1.1	.30
PV08	.37	.20	1.06	.8	1.11	1.1	.32
PV33	.33	.20	.78	-3.1	.75	-2.7	.62
PV06	.29	.20	.96	-.5	.93	-.7	.44
PV19	.29	.20	.90	-1.3	.86	-1.5	.50
PV23	.29	.20	.93	-.9	.89	-1.1	.47
PV40	.29	.20	.96	-.5	.97	-.3	.43
PV28	.25	.20	1.10	1.3	1.14	1.4	.28
PV03	.21	.20	.97	-.4	.94	-.6	.42
PV32	.09	.20	.83	-2.5	.78	-2.5	.57
PV44	.09	.20	1.29	3.7	1.42	4.1	.07
PV27	-.03	.20	.89	-1.5	.85	-1.7	.50
PV39	-.11	.20	1.00	.0	.96	-.4	.40
PV37	-.15	.20	1.15	2.1	1.17	1.8	.23
PV43	-.23	.20	.95	-.8	.92	-.9	.44
PV26	-.27	.20	.95	-.7	.98	-.2	.43
PV04	-.31	.20	.98	-.3	.87	-.2	.40
PV34	-.34	.20	.92	-1.2	.93	-1.4	.47
PV41	-.34	.20	.97	.4	.73	-.7	.41
PV11	-.42	.20	.80	-3.1	1.73	-3.0	.59
PV35	-.50	.20	1.38	4.9	.75	5.9	-.08

Table 4 (continued)

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
PV09	−.58	.20	.83	−2.6	.84	−2.5	.56
PV24	−.58	.20	.88	−1.7	1.08	−1.6	.50
PV18	−.67	.20	1.07	1.0	1.08	.7	.29
PV16	−.79	.20	1.18	2.2	1.25	1.9	.16
PV07	−.88	.21	.88	−1.5	.81	−1.6	.49
PV31	−1.01	.21	1.07	.9	1.09	.7	.27
PV15	−1.19	.22	.93	−.7	1.10	.7	.37
PV02	−1.23	.22	.88	−1.2	.77	−1.5	.47
PV10	−1.23	.22	1.01	.2	1.09	.6	.30
PV21	−1.43	.23	.97	−.2	.96	−.2	.34
PV14	−1.59	.23	.95	−.4	.92	−.3	.35
PV17	−1.64	.24	1.09	.7	1.23	1.1	.18
PV01	−1.76	.24	.90	−.7	.82	−.7	.40
PV42	−2.16	.27	1.01	.1	.87	−.4	.27

Note. PV = Phrasal Vocabulary Size Test item

from .20 to .28, which indicated that the item difficulty estimates were reasonably precise. Items PV46 (*to blame*), PV44 (*would appear*), and PV35 (*in effect*) underfit the model according to the infit MNSQ criterion, and items PV20 (*be expected to*) and PV11 (*so far*) underfit the model according to the outfit MNSQ criterion. Item PV35 had negative point-measure correlations, and PV 20 and PV46 have very low or no correlations, which are problematic because they indicated that more able test-takers missed the item and less able test-takers answered the item correctly. In other words, these items did not reliably distinguish between high and low ability test-takers.

An inspection of the distractor functioning revealed the reason for the negative point-measure correlation for item PV35 (*in effect*). It indicated that two distractors (*possibly* and *now*) were selected by 39% of the participants whose average person ability estimates were .00 and .31, respectively. The correct answer, *actually* was selected by 58% of the participants and their average person ability estimate was −.17, which was below that of the examinees who selected the two distractors.

An inspection of the distractor functioning of item PV20 (*be expected to*) also indicated problems as one distractor (*hoping to*) attracted 63% of the participants whose average person ability estimates was −.02. Only 15% of the participants selected the correct answer (*must*), and their average person ability estimates was −.11, which was lower than those of the examinees who selected the distractor.

An inspection of the distractor functioning of item PV46 (*to blame*) showed that 51% of the test-takers with average person ability estimates of .09 selected the distractor (*accusing anyone*). Only 24% of the participants selected the correct answer (*the cause of problem*), and their average person ability estimates was −.05, which was lower than those

of the test-takers who selected the distractor.

In order to examine whether items PV35, PV20, and PV46 disturbed the person measures, a Pearson correlation of the Rasch person ability estimates as estimated using all 50 items and the 47 items (excluding PV35, PV20, and PV46) was calculated. The Pearson correlation coefficient was r =.99, p < .001, which indicated that the item did not cause serious measurement problems; therefore, these 3 items were retained.

In sum, five out of 50 items (10%) misfit the Rasch model according to either of the infit or outfit MNSQ criteria, or negative value of point-measure correlations. Therefore, the majority of the items displayed good fit to the Rasch model. Even though the above items should be utilized with caution in the future, they are not problematic when viewed in the context of most of the 50 items as at least two other items have similar difficulty estimates as each of these items.

Research question 2 asked whether the Phrasal Vocabulary Size Test measures a single construct. To answer this question, the dimensionality of the items hypothesized to measure the same construct is investigated through a Rasch Principal Component Analysis of item residuals. 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) are 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.

The variance explained by the items (17.0%) was not greater than four times the variance accounted by the first contrast (4.5%). Therefore, the first criterion was not met. The Rasch model accounted for 25.7% of the total variance (eigenvalue = 17.3), which was below the required the value of 50%. The eigenvalue of the first residual contrast was 3.0, which was the same as the 3.0 criterion, so the third criterion was not met. The unexplained variance explained by first contrast (4.5%) 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.

Research question 3 asked how well and precisely each item in the Phrasal

Vocabulary Size Test measures test-takers' phrasal vocabulary knowledge. For this question, a Wright-map of the Phrasal Vocabulary Size Test is 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 1 shows the linear relationship between 101 test-takers and 60 items. On the far left side is the Rasch logit scale. Persons are indicated by the symbol '#' (representing two test-takers) . More able test-takers (i.e. higher-scoring persons) are toward the top of the figure and less able persons are toward the bottom.

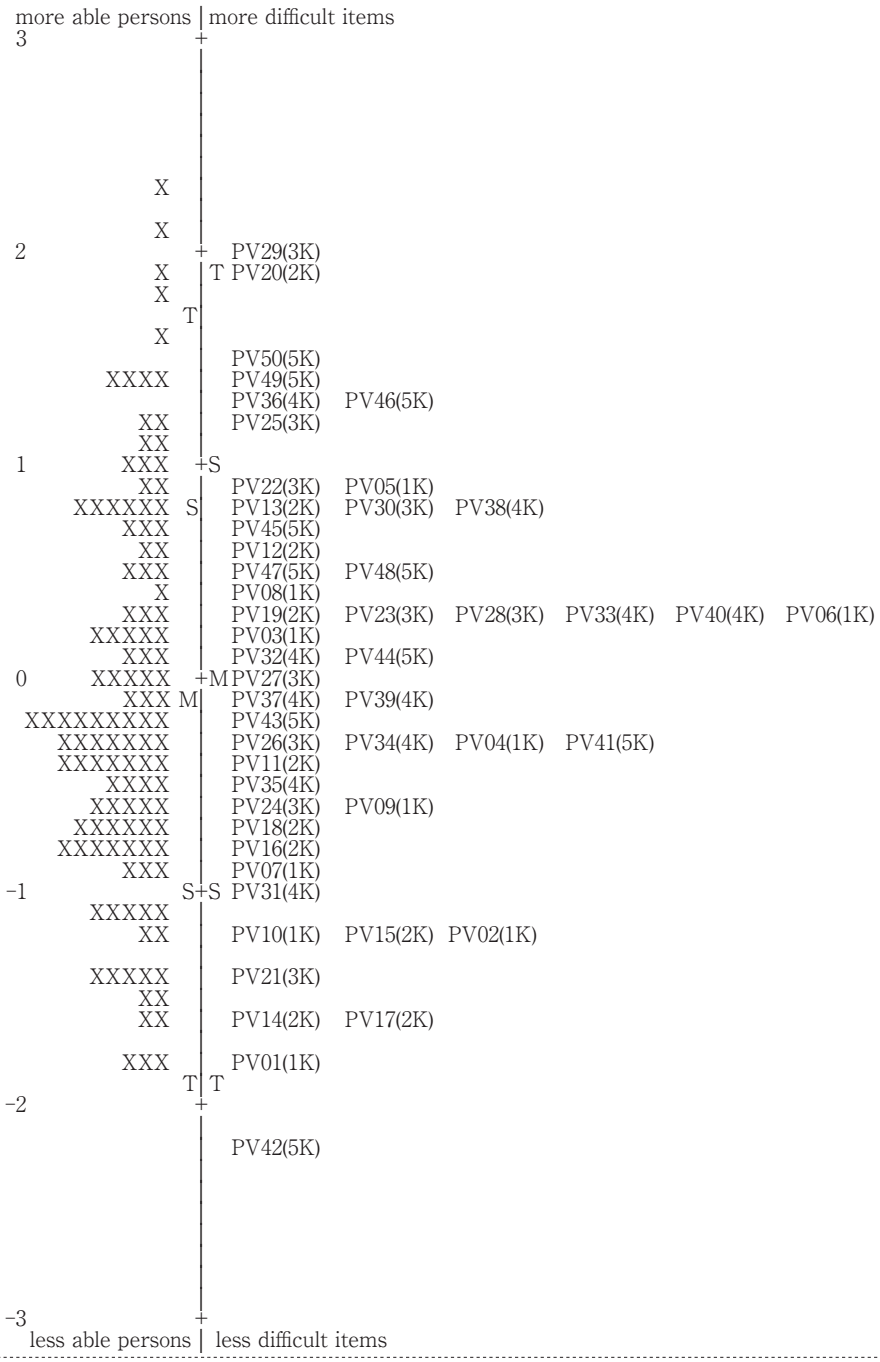
Figure 1 indicates that the Phrasal Vocabulary Size Test has a sufficient number of items, as the 50 items measure the full range of low and high-proficiency learners. No serious floor or ceiling effects were present for any examinees. The item mean is set to .00 (*SD* =.96) logits by convention, and the mean of the person ability estimates was −.11 (*SD* = .89). Also, there are no significant gaps in the empirical item hierarchy, as items are found along nearly the entire measurement range.

Research question 4 asked what the reliability and separation indices of the Phrasal Vocabulary Size Test 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 if these same items were given to another sample of comparable ability. 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) is adopted in this study. According to Fisher, person and item reliability of < .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.

The Rasch item reliability estimate was .95, which is excellent according to Fisher (2007), and the Rasch item separation index was 4.21, which indicated that items in Phrasal Vocabulary Size Test have at least four distinctive difficulty levels. The Rasch person

Figure 1. ***Wright map for the 50 items on Phrasal Vocabulary Size Test.***



Note. Each X equals 2 people; M = Mean; S = one standard deviation from the mean; T = two standard deviations from the mean. PV = Phrasal Vocabulary Size 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

reliability was .86, which was good according to Fisher (2007), and the separation index was 2.44, which meet the criterion.

Conclusion

This study was an initial validation of the Phrasal Vocabulary Size Test using Rasch Analysis. Overall, the result suggested that the vast majority of the PVST items fulfill the criteria of good measurement. That is, most items in the PVST adequately fit the Rasch model. No floor or ceiling effects were found. In addition, not only were there no serious gaps in the empirical item hierarchy, but there was also considerable redundancy, as items with similar difficulty estimates were found along nearly the entire measurement range. Moreover, they formed a fundamentally unidimensional construct, which implied that even though the PVST consists of grammatically different formulaic sequences, they are likely to measure a single construct. It is worth noting that this study was the first to investigate the unidimensionality of the items in the PVST.

One limitation of this study is that the participants are all female university students majoring in English. Therefore, the results of this study are only applicable to the similar participants. Future research can include male university students as participants to increase generalizability. In addition, future research needs to incorporate some qualitative approaches such as interview or thinking aloud to investigate test-takers' process of selecting correct answers including their test-taking strategies. These would help to ascertain test-takers' knowledge of target items and shed light on the following unanswered questions: to what degree does the format of multiple choice inflate the test-takers' knowledge of formulaic sequence, and why were test-takers with below average person estimates able to select the correct answers in item 20 (*be expected to*) or 46 (*to blame*), and why did test-takers with higher averaged person estimates miss these items?

References

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge University Press.

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions,* 21 (1), 1095. https://www.rasch.org/rmt/rmt211m.htm

Gyllstad, H., & Schmitt, N. (2019). *Testing formulaic sequence.* In Siyanova, A., & A. Pellicer (Eds.), Understanding formulaic language (pp. 174-191). Routledge.

Laufer, B. (1989). A factor of difficulty in vocabulary learning: deceptive transparency. *AILA Review,* 6, 10-20.

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

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

Martinez, R. (2011a). The development of a corpus-informed list of formulaic sequences for language pedagogy. [Unpublished doctoral dissertation]. University of Nottingham.

Martinez, R. (2011b). The Phrasal Vocabulary Size Test on Tom Cobb's LexTutor. Retrieved from http://www.lextutor.ca/tests/levels/reognition/phrase/

Martinez, R. & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics,* 33 (3), 299-320. https://doi.org/10.1093/applin/ams010

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

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher,* 31 (7), 9-13. https://jalt-publications.org/sites/default/files/pdf/the_language_teacher/06_2007tlt.pdf

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp.191-225). Longman.

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

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

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

Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics, 21* (4), 463-489. https//doi.org/10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wary, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.