A Validation Study of the Newly-developed Version of Vocabulary Size Test for Persian Learners

Seyyed Mohammad Reza Amirian,(Corresponding author)
Assistant Professor, Department of English Language, Faculty of Literature and Humanities
Hakim Sabzevari University, Sabzevar, Iran
Email: sm. amirian@hsu.ac.ir

Samaneh Salari
M.A student, Islamic Azad University, Neyshabur Science and Research Bracnh
Email: samanehsalari64@gmail.com

Zahra Heshmatifar
Ph.D. Student of TEFL,Department of English Language, Faculty of Literature and Humanities
Hakim Sabzevari University, Sabzevar, Iran
Email: heshmatifar50@gmail.com

Javad Rahimi
M.A. in Philosophy, Office of Education, Sabzevar, Iran
Email: rahimi2715@yahoo.com

Abstract
The present study addressed the development and the empirical validation of a recent bilingual Persian version of Vocabulary Size Test (VST). Moreover, an attempt was made to scrutinize the relationship between the previously-validated 14000bilingual version and this20000 version both of which measured receptive vocabulary knowledge. To accomplish this goal, the newly-developed version of VST was trialed on 294 Persian EFL learners from three proficiency levels, i.e. low, mid and high. The findings from Rasch Model analysis suggested that the new test functioned well because it fulfilled these criteria: it enjoyed a high level of reliability, it supported a unidimensional underlying construct, and the items formed a meaningful difficulty continuum. Furthermore, the results of one-way between-subject ANOVA indicated that the test was capable of distinguishing learners with different levels of proficiency. Moreover, it was revealed that the newversion was highly correlated with monolingual version as well as the previously-bilingual one. Finally, some implications calling for future research were reported.

Key words: Vocabulary Size Test, Validation, Persian, Reliability, Rasch Model.

1.Introduction

For many researchers(e.g. Knight, 1994; Laufer, 1997; Schmitt, 2000; Schmitt, 2008) vocabulary knowledge is at the very heart of all language skills in generaland of language comprehension in particular since the basic information that learners wish to comprehend is carried through lexical items (Nation, 2001).This ability to recognize a word and to recall its meaning when it is encountered is known as receptive knowledge which is connected with listening and reading, while

productive knowledge refers to the ability to produce a word when speaking or writing (Nation, 1990).Nation(2001)asserts that the performance of EFL/ESL readers, is largely affected by their knowledge of vocabulary. Accordingly, most of ESL/EFL readers complained about unfamiliar vocabulary as one of the huge obstacles in reading comprehension. What actually ESL/EFL readers concern about is not the inadequate practice of reading comprehension, but the lack of sufficient vocabulary knowledge (Haynes and Baker, 1993).Furthermore, Anderson and Freebody (1981) put forward a similar idea that the best predictor of understanding a text is vocabulary knowledge.

Apart from various aspects of vocabulary knowledge such as synonymy, antonymy, hyponymy, pronunciation, collocation, spelling and syntactic structure(Nation, 1990), two recent dimensions have been recognized as breadth and depth. The former refers to the quantity or the number of words that language learners know at a particular level of language proficiency, while the latter refers to the quality of their vocabulary knowledge (Bogaards& Laufer, 2004; Nation, 2001; Read, 2000). Admittedly, this distinction made in the field of vocabulary teaching and learning, has led into a revival of research as well as new insights into the nature of vocabulary assessment. (Bachman, 2000; Laufer, 1986).

Although both aspects of this two-dimensional approach are important, due to the vital role of vocabulary size as a successful predictor of academic achievement (Laufer & Goldstein, 2004; Nation & Meara, 2002; Saville-Troike, 1984), a renewed emphasis has been put on this issue by specialists in the field. As an example, Meara (1996) holds that the basic aspect of lexical knowledge is size. He argues

> "All other things being equal, learners with big vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies, and there is some evidence to support the view that vocabulary skills make a significant contribution to almost all aspects of L2 proficiency" (Meara 1996,p. 37).

Many researchers have made attempt to devise different tests tapping this specific facet of lexical knowledge i.e. vocabulary size such as the Eurocenters Vocabulary Size Test (Meara & Jones, 1990),Vocabulary Levels Test (VLT) (Nation, 1990, 2001; Schmitt, Schmitt, & Clapham, 2001;Beglar & Hunt, 1999)and Vocabulary Size Test (henceforthVST) (Nation & Beglar, 2007). It's worth mentioning that all these tests measure learners' ability to recognize a word in L2 and to recount one of its possible meanings in L1 i.e. receptive vocabulary knowledge.

VSTs are profitable for variety of reasons. They can be applied as placement or proficiency tests to group learners based on their vocabulary size and different proficiency level (Laufer & Nation, 1995). Following Beglar(2010),they can also serve as achievement or diagnostic tests to detect gaps in vocabulary knowledge of the learners and to keep a track of learners' growth in lexical proficiency and reading comprehension ability. Furthermore, they can be informative tools in the hands of curriculum designers to develop appropriate materials for the fulfillment of learners' needs.

In view of the mentioned purposes, currently, monolingual and bilingual versions of VSTs are being widely used in different ESL/EFL contexts.

1.1. Iranian Context: A Need for Monolingual or Bilingual VST?

According to Nation and Newton (2009), an ideal approach to language teaching and learning puts the same emphasis on all parts of language. For instance, learning a language in a communicative language teaching (CLT) context requires the mastery of four skills including listening, speaking, reading and writing as well as its components including grammar and vocabulary and

pronunciation. Therefore, within this framework the syllabus plan is designed in a way that all language skills and its sub-skills will be focused equally.

However, the findings of a number of research (Dahmardeh, 2009; Razmjoo & Riazi, 2006;Rahimi, 1996) revealed that in Iranian context where English is considered as a foreign language (EFL), the status of English Language Teaching (ELT) is no tin congruence with communicative program. In such a situation, the priority is mainly given to reading skill and vocabulary component. Rahimi's (1996) study on ELT methods taught in Iran, endorsed that Grammar Translation Method (GTM) has been the dominant approach to language teaching since 1950's. A more recent research conducted by Ghorbani (2009)also confirmed that in Iranian context, the central focus is put on reading skills to help learners read and translate English texts. Consequently, in order to satisfy the expectation of national curriculum which is achieving a high level of reading proficiency in English, learners are encouraged to translate the original English texts into their mother tongue (Persian) to comprehend it. This is why EFL learners are highly dependent on translation as a resort.

As it was noted, vocabulary size is closely associated with reading comprehension ability. In concrete terms, vocabulary size is, if not the best, a strong and critical predictor of the reading skill. (Nation, 2001, 2006; Qian, 1999, 2002). Hence, there's a dire need for a large vocabulary repertoire and subsequently for a useful assessment tool that measures total receptive vocabulary size of EFL learners as well as showing the extent of lexical gap they face in coping with reading materials.

For the purpose of the current study, the use of Persian bilingual VST is twofold. First, given the discussion above, in a reading-oriented context such as Iran, learners heavily rely on translation for effective comprehension. Accordingly, the aim of the bilingual test is to measure learners' recognition of word by providing L1 synonyms or definitions. Second, test takers can immediately comprehend the L1 definition or single-word equivalent much easier. Elgort's (2012) empirical study also confirmed this matter of simplicity when the bilingual version of the test resulted in higher scores. Similarly, Karami (2012) refers to the removal of long definitionsas one of the robust points in bilingual VST compared to the monolingual counterpart which callsfor good grammatical knowledge and high reading ability to detect the correct answer.

1.2. The previous and New Versions of VST

The foundation of the VST has been hinged upon this hypothesis that there is a close relationship between the word frequency and the probability that a word will be known. Findings of some empirical research (e.g. Hazenberg & Hulstijn, 1996;Meara & Jones, 1990) also supported the fitness of such a frequency-based pattern for most learners.

Because of the shortcomings of the Yes/No method of testing vocabulary size (Meara & Buxton, 1987), the multiple choice tests are widely used all around the world. In the former which is a checklist test, the learners are required to tick those words they know their meaning, whilein the latter presented in a four-choice format, the closest translation or definition of words is marked. One of the benefits of multiple-choice format for a vocabulary test compared to a rough Yes/No scale is encouraging learners to make use of their partial knowledge and intuitive guesses when they do not know the precise meaning of the word.

One of the most widely-used tests to estimate learners' vocabulary size is Beglar and Nation's(2007)VST as "a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1,000 to the 14th 1,000 word families in English" (Beglar & Nation,2007, p. 9). Unlike the VLT (Nation, 1990) which is basically a diagnostic test and tries to measure which wordlists learners should focus their efforts on, VST, on the other hand, is a proficiency measure for determining a learner's receptive vocabulary size(Beglar & Nation, 2007).

The 14000 version consists of 140 multiple-choice items, with 10 items from each 1000 word family level based on a frequency count of word families in the British National Corpus (BNC) (Leech, Rayson & Wilson, 2001). In order to get the total receptive vocabulary size, the total score needs to be multiplied by 100 since each item represents 100 word families. A more recent version has been designed by the same authors containing 100 multiple-choice items in which the total score needs to be multiplied by 200 to obtain the total receptive vocabulary size out of 20000. This new version is available in the form of two parallel and identical versions (versions A & B). Each item, on the VST, is presented in a decontextualized sentence, with four possible definitions of which one alternative is correct, and three are distracters. An example item of both bilingual and monolingual versions of VST is provided below:

dig: Our dog often <digs>.
a solves problems with things
b creates a hole in the ground
c wants to sleep
d enters the water

dig: Our dog often <digs>.
a. حل کردن
b. حفر کردن
c. خوابیدن
d. شنا کردن

The current study was an attempt to examine the validation ofa newly-developed bilingual version (version A)of VST and compare it with the monolingual version of the test and that of bilingual 14000 version as well.

2. Literature Review

2.1. Earlier Validation Studies

Traditionally, validity as a property of test itself was viewed as the extent to which a test measures what it is intended to measure (Kelly, 1927). By accepting this definition, validity was a divisible entity consisting of three types: criterion, content, and construct validity.

Current views of test validity tend to conceive of the concept of validity as a unified though multifaceted concept that focuses on inferences and consequences and values based on the test use. (Messick, 1989).Due to changes in the concept of validity which have been largely adopted in L2 research (Bachman, 2000),recent studies, especially in educational measurement, have made an attempt to collect empirical evidences for examining the intended or unintended outcomes of the test use as the validation evidence. Hence, some of the relevant empirical literature on empirical validity is reviewed below.

Beglar (2010) conducted a Rasch-based validation study of the monolingual VST among one-hundred-seventy Japanese and nineteen native English speakers. The results suggested that the test provides a satisfactory measure of test-learners' overall vocabulary size because of these features: a single-factor measurement, a high level of practicality in terms of both scoring and administration, unambiguous items, high reliability indices, and distinguishing different proficiency levels.

In another validation study on the vocabulary size of Vietnamese learners of English, Nguyen and Nation (2011)reported that bilingual versions of the VST can be more efficient than its time-consuming counterpart i.e. monolingual one. Furthermore, the authors, in contrast to prior accepted assumption, spoke of the necessity to sit all levels of the test. Nonetheless, in terms of distinguishing learners of different proficiency levels, the findings showed that the bilingual version works in much the same way as the monolingual test.

Irina Elgort's (2012) study was also an attempt to compare the English-Russian VST and the monolingual version using regression analysis. The results of the study suggested that bilingual format of the test is a more effective indicator of language learners' ability to recognize the right conceptual meaning of a test item. On the other hand, the analysis revealed that for high-level proficiency learners such as upper-intermediate or advanced learners, a monolingual test is likely to work efficiently as well. In other words, in favor of monolingual version for advanced learners, it has been argued that it may better reflect the construct of robust L2 word knowledge (Kroll & Stewart, 1994). Additionally, on the basis of another finding of the study, the behavior of cognates was found to be relatively the same between the monolingual and bilingual versions of the VST, and test takers outperformed in both test versions for the items involving cognates compared to non-cognates, especially in case of low-frequency words.

In a more recent experiment, Karami (2012) trialed a Persian bilingual version of the VST. He investigated the theoretical rationales as well as empirical analysis of validity argument based on the different aspects of Messickian (1989) validity. The findings of the study revealed that on the theoretical part, the test items seemed to be effectively sampled from a large corpus. The test also benefited from efficient administration, easy scoring and interpretation. On the empirical part, the study showed that the test effectively distinguished between test takers from different proficiency levels. In addition, the results of a factor analysis indicated that the test estimated a single construct, i.e. vocabulary knowledge.

3.Method

3.1.The Development of the Bilingual Persian Version

The first step in the process of the development of the bilingual Persian version was to translate the English alternatives into Persian. Effort was made to translate most of the alternatives into single-word equivalents, however, for some options, the possible equivalents were recognized to be phrases or complete sentences. Following Karami (2012), in terms of the translation of the cognates, L1 short definitions were used instead of the same loan word in the alternatives. It should be noted that the new 20000 version shared a number of items with the earlier 14000 version translated by Karami (2012).For this reason, the translation of the identical items were adopted from the pool of items available. After the primary draft had been translated by the two authors, it was given to four Persian reviewers each of whom left his or her comments for revisions or replacements of the vague translations. Three of the reviewers were post-graduate TEFL teachers with more than fifteen years of experience and one of them was an assistant professor who held a Ph.D. in TEFL. As expected, there were some opposing ideas regarding the substituting or removing some translations. Therefore, a panel discussion of the same reviewers was held. Finally, the translation of items were finalized according to the agreements of the panelists.

3.2.The Pilot Study

The next step before conducting the main study was to run a pilot study in order to check the effectiveness of the test. It was administered to twenty students enjoying the characteristics similar to the target group. In this phase of the study, the average time needed to take the newly-developed test and the ambiguous Persian equivalents were detected. For instance, it was found that some of the translations needed to be revised since they were vague or they signaled non-meaning clues for choosing the right answer such as the length of the choice.

3.3.The Main Study

3.3.1. Participants

A total of 294 Iranian male and female students with an age range of 17 to 30took part in this study (Table 1). These participants were all native Persian speakers. They were divided into three groups of elementary, intermediate and high proficiency based on their educational backgrounds. One-hundred-one of them studying the same English textbook as a required course in public high schools were considered as the low-proficient learners. The intermediate group contained eighty-nine junior and senior learners majoring English Literature at University of Sabzevar. Given that the admission to the collage at under-graduate level in Iran is done through a nationwide entrance examination, the English proficiency level of the under-graduate students could most probably be considered as similar. The high proficient group(N= 14) were either MA or Ph.D. students who majored in Teaching English as Foreign Language (TEFL), and all of them had at least six-year experience in studying English as their major. Therefore, it was expected that their proficiency level must be meaningfully higher than those of the mid group. Moreover, to ensure that all of the participants in each level of proficiency had learned English within the same context and were homogeneous in terms of their common experience in English learning, those who reported that they had simultaneously enrolled at private language institutes or any other English classes were excluded from the study.

Table1
Demographic Information of Participants

Proficiency level	Gender		Number of students
	Male	Female	
Low	83	108	191
Mid	8	81	89
High	6	8	14
Total	97	197	294

3.3.2. Data Analysis

Data including dichotomously scored results of the Persian bilingual version of VST were entered into anExcel data file which then were imported into WINSTEP software program version 3.72.3 in order to calculate statistics for the Rasch Model(Linacre, 2011). The generated output from Winsteps consisted of item- person map, separation reliability, item INFIT and OUTFIT statistics, and item difficulty map. Additionally, SPSS (version17.0) was used to calculate the one-way between-groups ANOVA.

3.3.3.Validation of the Test

The following questions guided the validation phase of the study:

1. Is the test capable of differentiating between learners from different proficiency groups?
2. Does the test benefit from a high level of reliability?
3. Is the item difficulty hierarchy meaningful along the test?
4. Does the test support a unidimensional underlying construct?
5. What is the relationship between the newly-developed 20000 bilingual version and the previously 14000 bilingual version?
6. What is the relationship between 20000 bilingual Persian VST and 20000 monolingual version?

4. Results

1- Is the test capable of distinguishing learners from different proficiency?

An evidence sought for statistical validity is whether the test distinguish between high and low abilities with sufficient statistical certainty. That is, more proficient examinees would significantly outperform the less proficient groups.

To compare the mean scores of three groups of test takers, one-way between-groups ANOVA was run. Table 2 indicated the descriptive statistics for the low, mid and high-proficiency groups. As the table displayed the mean scores of elementary group was 30.49, while middle group and high groups' mean scoreswere40.72 and 55.43, respectively.

Table 2
Descriptive Statistics

Group	N	Mean	Std. Deviation	Minimum	Maximum
Low	191	30.49	2.305	25	36
Mid	89	40.72	1.907	35	46
High	14	55.43	2.174	52	59

Figure 1 delineated the significantly different performance of three groups with the 95% confidence intervals for the means.

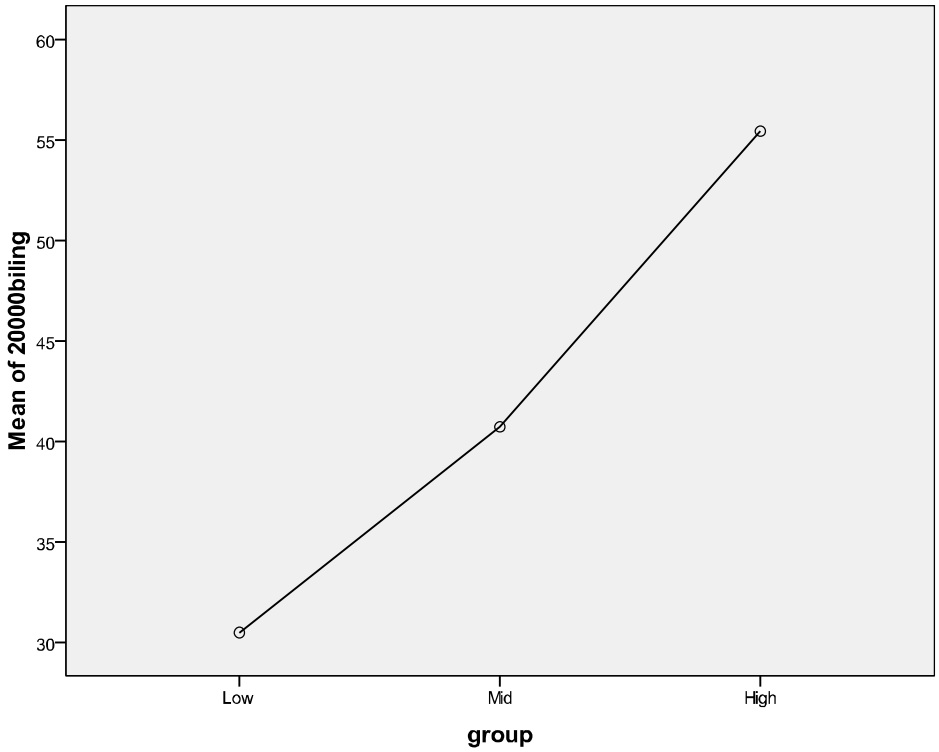

Figure 1. A Display of Groups' Performance

In order to ensure that the statistical test, i.e. one-way between-groups ANOVA fits the data, the homogeneity of variances was examined applying Levene's test. The Sig. value was.082 which is greater than alpha set at.05 level of significance. This supported the assumption that there was homogeneity of variances from one group to another.

Table 3displays the results of the one-way between-subjects ANOVA. As the table clearly presents, there is a significant difference between the means of at least two groups, F (2, 291) =1320.848, p<.05. Given that the desired large effect is greater than .14 (Cohen, 1988), the effect size of .58,using eta-squared statistic, was found to be large enough. In addition, to examine which mean differences were significant, the results were subjected to a post-hoc analysis. Tukey HSD was conducted to reveal a two by two comparison of three mean differences. The results obtained from this statistical test are presented in Table4.

Table 3
ANOVA Results

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12628.608	2	6314.304	132.848	.000
Within Groups	1391.123	291	4.780		
Total	14019.731	293			

Table 4
Multiple Comparisons

		Mean Difference			95% Confidence Interval	
(I) group	(J) group	(I-J)	Std.Error	Sig	Lower Bound	Upper Bound
Low	Mid	-10.232*	.281	.000	-10.89	-9.57
	High	-24.942*	.605	.000	-26.37	-23.52
Mid	Low	10.232*	.281	.000	9.57	10.89
	High	-14.709*	.629	.000	-16.19	-13.23
High	Low	24.942*	.605	.000	23.52	26.37
	Mid	14.709*	.629	.000	13.23	16.19

*. The mean difference is significant at the 0.05 level.

Table 4 shows the results from multiple comparisons. It is evident from the inspection of the P-values that all the mean differences were significant at the.05 level. In other words, the mean performances of the three groups were significantly different from each other.

By and large, the results of the one-way between- subjects ANOVA indicated that the test has the power to effectively distinguish the groups with different proficiency levels.

2-Does the test benefit from a high level of reliability?

To answer this research question, the computation of reliability index is done under the Rash Model. The rationale of using Rasch Model analysis is that in reliability the consistency and reproducibility of persons' locations are of more importance than the reproducibility of items(Baghayi, 2009). Compared to the classical test theory in which reliability was mainly estimated through the statistics of the items such as the classical Kuder-Richardson 20 (KR-20) or Cronbach's alpha, the Rasch model generates separation reliability to show how reliably the persons are separated on the ability continuum.

Andrichand Marasis (2006) found that although KR-20 and Rasch separation reliability have many features in common, the former suffers from the problems caused by small groups and missing data. Irrespective of such differences, Baghayi(2009) refers to person reliability as an equivalence to the classical test reliability.

Table 5 provides the summary of item and person information. The first row in the bottom of the table is called 'Real' section which gives the observed statistics and the second row is 'Model' which indicates how the condition would have been under the ideal situations that data fit the Rasch model.

Table 5
Rasch Summary of 294 measured person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	34.8	100.0	-1.35	.43	.98	-.1	.92	.0
S.D.	6.9	.0	1.20	.02	.27	1.1	.85	.8
MAX.	59.0	100.0	2.52	.45	1.75	2.9	9.90	5.0
MIN.	25.0	100.0	-3.21	.38	.38	-2.8	.15	-1.3

REAL RMSE .45 TRUE SD 1.12 SEPARATION 2.47 PERSON RELIABILITY .86
MODEL RMSE .43 TRUE SD 1.12 SEPARATION 2.61 PERSON RELIABILITY .87
S.E. OF PERSON MEAN = .07

As the results in Table 5 reveals, the estimated person reliability is .86 indicating that the reliability index along the test is high. That is, most probably persons (or items) estimated with high measures actually do have higher measure than persons (or items) estimated with low measures. Thus, through the Rasch scaling analysis, it may be safely concluded that the test enjoys a high level of reliability.

3-Is the item difficulty hierarchy meaningful?

The hypothesis that the items in a test have been meaningfully distributed along it, supports the substantive component of construct validity. In this kind of validation, the items are selected based on the empirical confirmation. Thus, the inclusion and exclusion of items should be informed by the statistical features of the items. The Rasch model was used to examine this aspect of validity because it offers a detailed inspection of the spread of items along the test. In fact, in Rasch analysis the item hierarchy that is created by the item difficulty estimates provides an indication of construct validity (Smith, 2001).The Rasch measure is calculated with the following formula:

$$\log [pni/(1- pni)]= Bn- Di(1)$$

*Bn*is the ability of a person n and *Di*is the difficulty of item *i*.

Furthermore, the Rasch Model makes it possible to map the results linearly since it transforms raw item difficulties and raw person scores into equal interval measures. Item difficulty is described on a continuum from less difficult to more difficult and is calibrated in logits. A logit is a unit of measurement used in Rasch analysis for calibrating items and measuring persons, based on the natural logarithmic odds of the probability of an answer. The validity of interpreting items in terms of difficulty are verified by investigating whether all the items are hierarchically ordered to define the construct.

Table 6reveals the spread of item-difficulties and person-abilities. The most difficult items are shown at the top of the map and on the right of the border while the most capable students are on the highest left side of the model. To have a 50% chance of responding to an item correctly, a student has to be plotted at the same level as the item. For the items above the person, it means that there is less than a 50% chance of answering that item correctly, and these items are estimated to be beyond the students' ability level.

Since items which are too easy for test-takers are unlikely to detect differences among persons with different levels, it is evident in the map that difficulty locations of items are above the ability of students and they are widely dispersed. In other words, the pool of persons is located below the region where most of the items are located. Psychometrically speaking, whereas there are very few students with abilities over the 5 logits, several items falls at about 6 logits which is clearly above their abilities. In addition, the map indicates that no noticeable gaps (i.e., >.30 logits) could

be found across neither items nor abilities (See Table 6).This points to the fact that most area of the construct domain has been covered by the test.

Table 6
Map of Person and Item

```
------------------------------------------------------------------------------

       MAP OF PERSON AND ITEM
MEASURE                                  |                            MEASURE
<more> --------------------- PERSON  -+- ITEM    ----------------- <rare>
    6                                    +  XXXXXXXXXXXXXXXXXXXX             6
                                         |
                                         |
|  XXXX
    5                                    +                                    5
                                         |
|  XXXX
|  XXX
    4                                    +                                    4
|  XXXX
                                         |S
|  X
    3                                    +                                    3
|  X
                                    .    |  XXXXX
.##  |  XXX
    2                               ##   +  X                                 2
#  |  XX
#  |  X
                                         |
    1                                   T+  XXX                              1
|  XX
                                    ##   |  X
                                  ####   |  XXX
    0                          .######   +M XXXX                             0
.####################### S|  X
.######  |  XX
.###  |  X
   -1                                ##   +  X                               -1
                                     # M|
                        .###############  |
                         ###############  |
   -2                   .###############  +  XX                              -2
                      .#################  |
.################### S|  XXX
                              ########   |
   -3                              .##   +  XXX                              -3
                                     .   |  X
                                         |S X
                                       T|  X
   -4                                  +  X                                 -4
|  XXXX
                                         |
                                         |
   -5                                    +  XX                              -5
|  XXX
|  X
                                         |
   -6                                    +  X                               -6
                                         |
|  X
|  X
   -7                                    +T XXXXXXXX                        -7
<less> --------------------- PERSON  -+- ITEM    ----------------- <frequent>
 EACH "#" IN THE PERSON COLUMN IS 2 PERSON: EACH "." IS 1
```

The presence of difficult items at the high end of the difficulty spectrum leads to the prevention of the ceiling effect. In General, the results suggest that the difficulty hierarchy of the test is in agreement with the hypothesized model of construct validity. In fact, not only the

distribution of items can effectively detect changes in test-takers, but also it well-targeted for the sample under study.

4-Does the test estimate a single construct?

One of the ways of demonstrating construct validity of a test is to establish unidimensionality. This assumption implies that a test should measure one single dimension or construct at a time. Due to the fact that no test is purely unidimensional in the real world (Wright & Stone, 1979), the new conceptualization suggests that it is not an absolute matter but it is a matter of degree. Rasch compensates for this by using interval scaling. Therefore, unidimensonality in Rasch Model is psychometric dimensionality rather than psychological one. It means "a single underlying measurement dimension; loosely, a single pattern of scores in the data matrix" rather than "a single underlying (psychological) construct or trait" (MacNamara, 1996, p. 271).In fact, unidimensionality is ascertained if the data fits the Rasch Model.

To answer this research question,fit statistics were examined to check how well the items define the unidimensional construct. Misfitting items are indications of violations of the unidimensionality principle and they are examples of multidimensionality and candidates for modification, discard or according to Bond and Fox (2007), clues to think about amending our construct validity.

Table 7 indicates the fit indices for some of the items. The items are arranged from difficult to easy. The first column, "ENTRY Number", shows the number given to each item in the test ranging from 1 to 100. The second column, labeled as "TOTAL SCORE", represents the total score for each item i.e. the number of participants who have responded correctly to that item. Third column labeled as "COUNT" indicates the number of participants who have attempted each item. Next column labeled as "MEASURE" gives the difficulty estimates for the items. In the fifth column, "MODEL S. E." displays the standard error of the item difficulty measures. "MNSQ" and "ZSTD" are abbreviations for "mean-square" and "z standardized distribution" respectively, and are provided for "INFIT" as well as "OUTFIT" columns.

Table 7

Item statistics: Measure order

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
64	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item64
67	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item67
70	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item70
72	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item72
73	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item73
74	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item74
75	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item75
77	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item77
80	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item80
83	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item83
85	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item85
86	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item86
88	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item88
89	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item89
90	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item90
92	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item92
95	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item95
96	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item96
98	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item98
100	0	294	6.52	1.83		MAXIMUM MEASURE			.00	.00	100.0	100.0	Item100
65	1	294	5.30	1.01	.95	.3	.78	-1.5	.17	.10	99.7	99.7	Item65
78	1	294	5.30	1.01	.95	.3	.78	-1.5	.17	.10	99.7	99.7	Item78
81	1	294	5.30	1.01	.95	.3	.78	-1.5	.17	.10	99.7	99.7	Item81
84	1	294	5.30	1.01	.92	.2	.76	-1.7	.19	.10	99.7	99.7	Item84
62	2	294	4.59	.72	.92	.1	.81	-1.2	.23	.14	99.3	99.3	Item62
66	2	294	4.59	.72	.93	.1	.79	-1.3	.24	.14	99.3	99.3	Item66
87	2	294	4.59	.72	.96	.2	.82	-1.1	.21	.14	99.3	99.3	Item87
91	2	294	4.59	.72	.93	.1	.92	-1.3	.24	.14	99.3	99.3	Item91
76	3	294	4.16	.59	.93	.0	.71	-1.0	.27	.17	99.0	99.0	Item76
82	3	294	4.16	.59	.98	.1	.80	-1.5	.23	.17	99.0	99.0	Item82
99	3	294	4.16	.59	.95	.1	.79	-1.6	.25	.17	99.0	99.0	Item99
61	4	294	3.85	.52	1.05	.3	1.33	.7	.15	.20	98.6	98.6	Item61

68	4	294	3.85	.52	.86	-.2	.78	-1.1	.30	.20	98.6	98.6	Item68
71	4	294	3.85	.52	.83	-.2	.79	-2.0	.34	.20	98.6	98.6	Item71
79	4	294	3.85	.52	.86	-.2	.70	-1.9	.33	.20	98.6	98.6	Item79
63	7	294	3.23	.40	.75	-.7	.80	-2.0	.44	.25	97.6	97.6	Item63
58	10	294	2.81	.34	1.15	.6	1.36	.8	.21	.29	96.6	96.6	Item58
93	12	294	2.59	.32	.58	-1.9	.12	-1.7	.56	.32	95.9	95.9	Item93
36	13	294	2.49	.31	1.31	1.3	1.30	2.0	.06	.32	95.2	95.6	Item36
54	13	294	2.49	.31	1.18	.8	1.34	1.5	.16	.32	95.2	95.6	Item54
97	13	294	2.49	.31	.78	-2.0	.92	-2.0	.58	.32	95.9	95.6	Item97
56	14	294	2.40	.30	.75	-2.0	.72	-.7	.55	.33	95.6	95.2	Item56
69	15	294	2.32	.29	.81	-1.8	.82	-1.1	.63	.34	95.6	94.9	Item69
94	16	294	2.23	.28	.79	-1.6	.88	-.1	.49	.35	94.6	94.6	Item94
34	17	294	2.16	.27	1.26	1.3	1.33	2.8	.11	.36	92.9	94.3	Item34
44	21	294	1.88	.25	1.26	1.4	1.71	1.7	.20	.38	91.5	93.3	Item44
50	22	294	1.82	.25	1.27	1.5	1.07	2.0	.16	.38	91.2	93.0	Item50
38	24	294	1.70	.24	1.30	1.9	1.31	1.3	.19	.39	89.5	92.6	Item38
52	27	294	1.54	.23	.97	-.1	1.01	.1	.40	.41	92.2	91.8	Item52
35	38	294	1.05	.20	1.27	2.0	1.32	1.5	.22	.44	83.0	89.1	Item35
43	38	294	1.05	.20	1.28	2.0	1.21	1.9	.10	.44	87.1	89.1	Item43
48	41	294	.94	.19	.98	-.1	1.19	.9	.42	.44	90.1	88.3	Item48
17	47	294	.73	.18	.84	-1.5	.79	-1.1	.56	.46	88.8	86.6	Item17
53	48	294	.70	.18	.90	-.9	1.05	.3	.50	.46	88.4	86.3	Item53
16	57	294	.43	.17	1.07	.7	1.14	.8	.41	.47	83.3	83.7	Item16
46	60	294	.34	.17	1.27	1.4	1.01	1.2	.09	.47	81.0	82.9	Item46
18	62	294	.29	.16	.99	-.1	1.05	.3	.47	.47	83.7	82.3	Item18
26	62	294	.29	.16	.90	-1.1	.73	-2.0	.58	.47	79.6	82.3	Item26
23	76	294	-.06	.15	.89	-1.4	.81	-1.5	.57	.48	79.3	78.5	Item23
39	76	294	-.06	.15	1.27	1.5	1.19	1.1	.07	.48	69.0	78.5	Item39
51	76	294	-.06	.15	.75	-1.9	.75	-1.5	.75	.48	82.0	78.5	Item51
27	77	294	-.08	.15	.72	-1.8	.77	-1.0	.69	.48	82.3	78.3	Item27
59	90	294	-.37	.15	.77	-1.9	.85	-1.9	.72	.49	87.1	76.1	Item59
60	93	294	-.43	.14	.79	-1.8	.78	-1.1	.78	.49	89.5	75.8	Item60
57	101	294	-.59	.14	.88	-1.8	.84	-1.7	.57	.49	80.3	75.1	Item57
24	109	294	-.75	.14	.74	-1.4	.98	-1.1	.80	.48	93.9	74.5	Item24
47	128	294	-1.10	.13	1.16	1.9	1.43	2.0	.34	.47	68.0	72.8	Item47
33	173	294	-1.88	.13	.97	-.5	.93	-.6	.45	.42	72.1	68.8	Item33
31	174	294	-1.90	.13	1.12	1.2	1.29	1.4	.28	.42	69.0	68.8	Item31
40	206	294	-2.48	.14	.98	-.4	.89	-.7	.39	.37	72.8	72.0	Item40
42	209	294	-2.54	.14	.92	-1.5	.77	-1.5	.44	.36	75.2	72.6	Item42
25	210	294	-2.56	.14	1.13	1.2	1.25	1.5	.24	.36	71.4	72.8	Item25
30	226	294	-2.89	.15	.92	-1.1	.73	-1.5	.41	.33	75.5	77.0	Item30
7	229	294	-2.95	.15	.98	-.2	.85	-.7	.35	.32	77.2	78.0	Item7
21	234	294	-3.07	.15	.96	-.5	.72	-1.4	.37	.31	79.3	79.6	Item21
49	241	294	-3.24	.16	1.17	1.9	1.85	2.0	.08	.29	82.0	82.0	Item49
37	247	294	-3.40	.17	.97	-.3	.98	.0	.29	.27	84.0	84.0	Item37
29	257	294	-3.70	.18	.96	-.3	.69	-1.1	.30	.24	87.4	87.4	Item29
15	266	294	-4.04	.20	1.04	.3	.86	-.4	.19	.21	90.5	90.5	Item15
8	269	294	-4.17	.21	1.06	.4	1.21	.7	.14	.20	91.5	91.5	Item8
22	270	294	-4.22	.22	1.04	.3	1.26	.9	.14	.20	91.8	91.8	Item22
10	272	294	-4.31	.23	.99	.0	.73	-.8	.21	.19	92.5	92.5	Item10
20	272	294	-4.31	.23	1.14	.8	1.20	1.4	.12	.19	92.5	92.5	Item20
55	281	294	-4.89	.29	1.00	.1	1.35	1.3	.11	.15	95.6	95.6	Item55
32	282	294	-4.98	.30	.98	.0	.73	-.6	.17	.14	95.9	95.9	Item32
12	284	294	-5.17	.33	.99	.1	.75	-.7	.16	.13	96.6	96.6	Item12
19	285	294	-5.28	.34	.97	.0	.79	-.9	.17	.12	96.9	96.9	Item19
41	285	294	-5.28	.34	.99	.1	.82	-.8	.15	.12	96.9	96.9	Item41
13	287	294	-5.55	.39	1.02	.2	.90	.0	.09	.11	97.6	97.6	Item13
14	289	294	-5.90	.45	.99	.1	.92	-.8	.13	.09	98.3	98.3	Item14
45	291	294	-6.42	.58	.99	.2	.72	-.7	.11	.07	99.0	99.0	Item45
4	292	294	-6.83	.71	.98	.2	.77	-1.1	.11	.06	99.3	99.3	Item4
1	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item1
2	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item2
3	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item3
5	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item5
6	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item6
9	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item9
11	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item11
28	294	294	-8.73	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	Item28

The acceptable range for the standardized infit and outfit values ("ZSTD") is between -2 and +2, and for mean square values ("MNSQ") between 0.70 and 1.3. Values greater than 1.3 show significant misfit, i.e. lack of predictability, while values below 0.7 show significant overfit, i.e. less variation than might normally be expected(Linacre, 2007).

Table 7 shows that the infit indices are within the acceptable range. It means that the items are working harmoniously to define the variable. Although the mean square values for outfit is a little bit out of the range, the ZSTD is within the acceptable range, therefore it does not cause a serious problem. Items 64, 67, 70, 72, 73, 74, 75, 77, 80, 83, 85, 86, 88, 89, 90, 92, 95, 96, 98, 100 are the most difficult items on the test. From 100 participants who have attempted these items, no one could get them right. The difficulty of these items are estimated to be 6.52 logits with the standard error of 1.83. This means one can be 95% sure that the true value for the difficulty of these items lies somewhere between 2.86 to 10.18 logits, i.e., two SE's below and above the observed measure.

The "PT-MEASURE" column indicates the observed ("CORR.) as well as the expected ("EXP.) correlation between performance on each item and the ability measures of the participants who have answered that item correctly. As it is evident in Table 7,positive item measure correlations suggest that the items are in the same direction of measurement. The correlation index for items 1,2,3,5,6,9,11,28 is zero. This indicates that these items are too easy for the participants i.e., below the lowest person ability logit score. Thus, these items do not contribute to the discrimination of person abilities because all test-takers answered them correctly. Additionally, there are several items that are above person abilities, i.e., items 64,67, 70, 72, 73,74,75, 77, 80, 83,85, 86, 88, 89, 90, 92,95, 96, 98, 100.These items may distort the measurement(Wolfe & Smith, 2007) and they are targets for inspection, reconsideration or discarding(McNamara, 1996).Nevertheless, it is suggested to keep the extremely difficult items as well as the extremely easy ones because of two reasons: 1)the test has been developed to estimate vocabulary size of the participants with different proficiency levels; 2) the selection of items is on the basis of frequency level, hence the existence of cognates as well as the incorporation of most or least frequent words is naturally inevitable.

In short, on the basis of fit analysis, one may conclude that the deviation of observed data from the model's prediction is reasonably tolerable. That is, for the purpose of this study, the data are in accordance with the fitted Rasch model. In other words, little evidence was found to highlight the construct-irrelevant variance in the test and the items are assessing the single underlying construct i.e. receptive vocabulary knowledge.

5-What is the relationship between20000bilingual Persian VSTand that of 14000 version?

To investigate the relationship between the newly-developed 20000 bilingual test with the 14000 English-Persian version, Pearson product-moment correlation was calculated. It is worth pointing out that the previously-developed bilingual version was administered two weeks later. Following similar studies (e.g. Harris, 1969), a two week span was chosen because less than this time the students might use their short term memory to answer the questions. A look at Table 8 indicates the mean difference between the earlier and the new bilingual versions. In order to find out whether the difference was statistically significant, Pearson correlation coefficient was run. Table 9 reveals that the Pearson correlation coefficient turned out to be.94which was statistically significant at $p < .01$.

Table 8
Descriptive Statistics

	Mean	Std. Deviation	N
20000 biling	34.7721	6.91729	294
14000 biling	60.5782	8.28826	294

Table 9
Correlation Results

		20000biling	140000biling
20000biling	Pearson Correlation	1	.948**
	Sig. (2-tailed)		.000
	N	294	294
140000biling	Pearson Correlation	.948**	1
	Sig. (2-tailed)	.000	
	N	294	294

**. Correlation is significant at the 0.01 level (2-tailed).

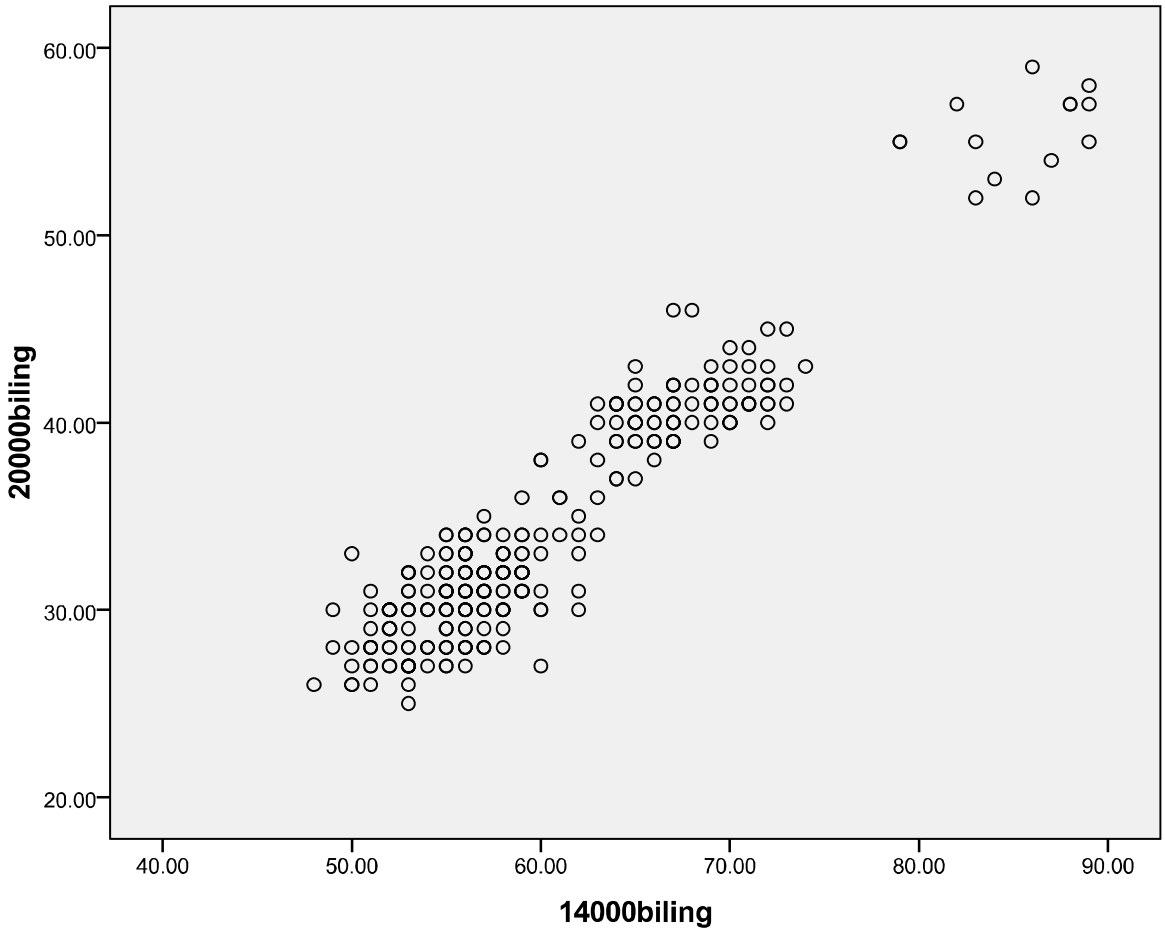

*Figure 2.*The Scatterplot of Tests' Relationship

Figure 2 displays the relationship between two tests. It is clear that the line graph brings out a positive relationship between two tests. Therefore, based on the results, it can be concluded that there exists a significant relationship between these two versions of VST. In other words, those who got higher score in the former version enjoyed higher score in the new version, too. Furthermore, the strong correlation coefficient between the two test scores also suggests that the two versions of the test were assessing the same construct, i.e., receptive vocabulary knowledge.

6-What is the relationship between 20000bilingual Persian VST and 20000 monolingual version?

To answer this research question, a second correlation coefficient was estimated to investigate the relationship between the newly-developed 20000 bilingual tests with the 20000 monolingual version. The descriptive statistics and the results of the correlations are presented in Tables 10 and 11. As it is evident there was a significant correlation between two versions at p <.01. Therefore, there was a close relationship between the learners' scores on 20000 monoloingual and bilingual VSTs (r=.96). In other words, as expected, those with a higher scores in monolingual version outperformed in bilingual version as well.

Table 10
Descriptive Statistics

	Mean	Std. Deviation	N
20000 monoling	32.9218	5.63909	294
20000biling	34.7721	6.91729	294

Table 11
Correlation Results

		20000monoling	20000biling
20000monoling	Pearson Correlation	1	.960**
	Sig. (2-tailed)		.000
	N	294	294
20000biling	Pearson Correlation	.960**	1
	Sig. (2-tailed)	.000	
	N	294	294

**. Correlation is significant at the 0.01 level (2-tailed).

To illustrate the relationship between two tests graphically, the scatter plot of tests' relationship was depicted. Figure 3 can very well demonstrates how the 20000 monolingual and bilingual versions are highly correlated.

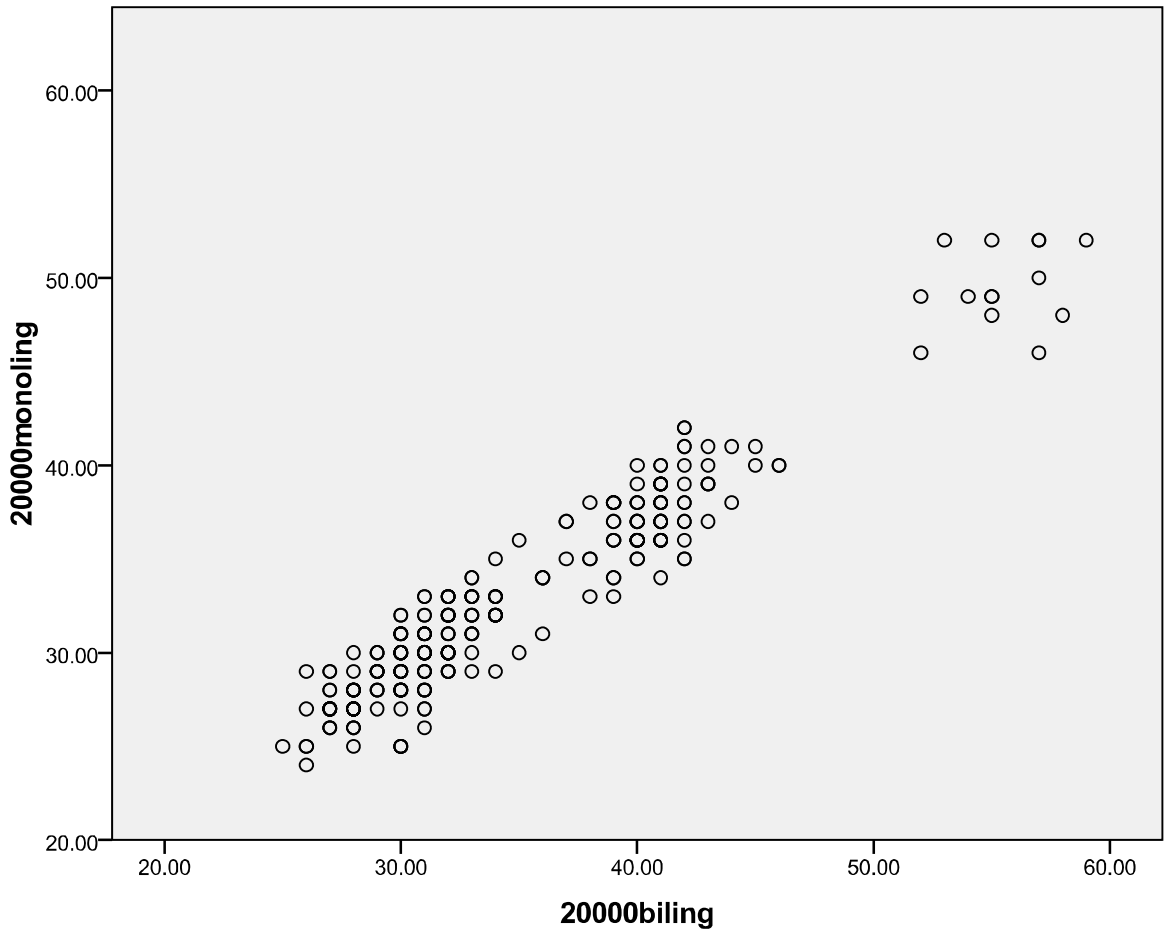

*Figure 3.*The Scatterplot of Tests' Relationship

5. Discussion

The results of different analyses provided validity evidence within Rasch framework since recent views tend to conceive of test validation as an ongoing process supported by empirical evidence (Messick, 1989).

It was found through Rasch scaling analysis that the test enjoyed a high level of reliability as a vital component of validity argument (Messick, 1995). This finding is consistent with the findings of Karami (2012)who found that the 14000 bilingual version of the test benefited from high reliability indices. In addition, based on the obtained results from fit statistics, the vast majority of items were assessing a single underlying factor. Therefore, the unidimensionality principle was confirmed. This is in line with the findings of Beglar (2010) who found that the monolingual version of the test was very clearly measuring a single construct i.e. written receptive vocabulary knowledge, while other factors did not play a major role in performance on the text. Moreover, in this study, map of person and item was used to check the representativeness of the items. There were no noticeable gaps in the item difficulty hierarchy. This pointed to the fact that the spread of item-difficulties and person-abilities was meaningful along the test. Taking into account the findings of Begalr's (2010) study in which the empirical item hierarchy showed reasonable spread, in this case, this finding is congruent with his finding.

With regard to the relationship between the earlier 14000 and the new 20000 versions, correlational statistics demonstrated a significant positive relationship between two versions. The same story was true about the relationship between 20000 monolingual and 20000 bilingual versions which were highly correlated. All in all, contrary to our expectation and against the common assumption in the literature(e.g. Elgort, 2012) that the bilingual version is far sensitive measure of written receptive vocabulary knowledge, this finding suggests that they can be safely used interchangeably taking different contexts and purposes into account.

Furthermore, following Messick (1995), an important aspect of validity evidence, is to study the differences in the performance of examinees from different levels.The results provided helpful evidence that the test was capable of distinguishing learners with different proficiency levels. Similar finding was reported by Nguyen and Nation (2010) andKarami (2012) that the test is capable to effectively distinguish between examinees from different proficiency levels. It means that the High group outperformed the Mid group which in turn scored higher than the Low group.

Another major finding revealed in this study was that different versions of VSTs were highly practical. The same finding has been pointed out by a number of researchers (e.g. Beglar, 2010; Karami, 2012) that both administration and scoring is done so easily. Indeed, as Nguyen and Nation (2011) rightly pointed out, if used appropriately, they can become a very useful tool for L2 research.

6. Conclusions and Implications

The primary aim of the present study was to provide some empirical analysis for validation of a newly-developed bilingual Persian version of the VST. It should be noted that With respect to Iranian EFL context where reading comprehension is considered as a framework within which other skills are defined (Anani Sarab, 2006), it seems essential to investigate the validation of a test estimating how many words learners know in a foreign language.

In general, the newly-developed Persian bilingual version of VST has been shown to work satisfactorily.It had very few misfitting items and it did exhibit reasonably well spread across items. By and large, the test is deemed to meet different Messickian construct-validity issues such as a high level of reliability, a unidimensional underlying factor, a meaningful difficulty hierarchy, and the capability of distinguishing learners with different levels of proficiency. Furthermore, it was revealed that there was a close correlation between 20000bilingualversion and those of 20000 monolingual version as well as 14000 bilingual one.

The present study, while attempting to provide validity evidence, did not include such analyses as differential item functioning (DIF) or item distractor analyses.These should be explored in more detail in the future to determine if there are any items that are causing unexpected response patterns either across groups or across sections of the test. The current study also leaves room for conducting longitudinal studies to track learner's progress in increasing vocabulary size over time. Another good topic for scrutiny might be the effect of guessing on the test takers' performance on the VST. Furthermore, following Beglar (2010), future studies could examine whether estimating vocabulary size by multiplying the obtained score on a 100-item test by 200 is an appropriate approach to interpreting test takers' vocabulary size.

By implication, this study outlines Rasch model as a systematic methodology for empirical validation studies which can be applied by language testing specialists to provide validity evidence for their tests. In addition, the newly-developed VST can be confidently used for needs analysis by Persian teachers and subsequently by other decision makers as likely the best indicator of their vocabulary size.

Acknowledgements

The authors would like to acknowledge Professor Paul Nation for his kind permission to use the recent monolingual versions of VST. Sincere thanks also goes to all the participants for generously sitting long hours to answer the tests.

References

Anani Sarab, M. R. (2006). *The Iranian curriculum for designing secondary school's English language textbooks.* Tehran.

Anderson R.C., & Freebody, P.(1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and Teaching: Research Perspectives*. Newark, DE: International Reading Association.

Andrich, D., & Marasis, I. (2006). *Instrument DesignWith Rasch IRT and Data Analysis1*. Unit Materials, Semester 2, 2006, Perth, Australia: Murdoch University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17* (1), 1-42.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction, 22*(1), 1145-1146. RetrievedApril 26,2014, fromhttp://www.rasch.org/rmt/rmt221a.htm.

Baghayi, P. (2009). *Understanding the Rasch Model.*Missaq Publishing Company, Mashad.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing 27*(1), 101-118.

Beglar, D.,& Hunt, A. (1999). Revising and validating the 2000 word level and university word level tests. *Language Testing 16*, 131-62.

Bogaards, P., & Laufer, B. (2004). *Vocabulary in a Second Language.* John Benjamins Publishing Company, Amsterdam.

Bond T.G., & Fox,C.M.(2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Lawrence Erlbaum.

Cohen, J.W. (1988).*Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: LawrenceErlbaum Associates.

Dahmardeh, M. (2009). *English language teaching in Iran and communicative language teaching*. A thesis submitted for the degree of PhD at the University of Warwick. Retrieved March 16, 2014, from http://go.warwick.ac.uk/wrap/2748

Elgort, I. (2012). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test.*Language Testing 30*(2), 253–272.

Ghorbani, M. R. (2009). ELT in Iranian high schools in Iran, Malaysia and Japan: Reflections on how tests influence use of prescribed textbooks. *Reflections on English Language Teaching, 8* (2), 131–139.

Harris, D.P. (1969). *Testing English asa second language.* New York, Mcc Graw-Hill Book Company.

Haynes, M., & Baker, I. (1993). American and Chinese readers learning from lexical familiarization in English text. In T. Huckin, M. Haynes, & J. Coady (Eds.), *second language reading and vocabulary learning* (pp. 130-152). Norwood, NJ: Ablex.

Hazenberg, S., & Hulstijn, J.H. (1996). Defining a minimal receptive secondlanguage vocabulary for non-native university students: an empirical investigation. *Applied Linguistics 17*, 145-163.

Karami, H. (2012). The Development and Validation of a Bilingual Version of the Vocabulary Size Test. *RELC Journal, 43*(1) 53 –67.

Kelley T.L. (1927). *Interpretation of educational measurements*. Yonkers, NY, World Book Company.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal, 78*(3), 285-299.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evi dence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*, 149–174.

Laufer, B. (1986). Possible changes in attitude towards vocabulary acquisition research. *IRAL, XXIV(1),* 69-75.

Laufer, B.(1997). The lexical plight in second language reading. In: Coady, J., Huckin, T. (Eds.), *SecondLanguage Vocabulary Acquisition.* Cambridge University Press, Cambridge.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computeradaptiveness. *Language Learning, 54*(3), 399-436.

Laufer, B.,& Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16* (3), 307- 322.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*.

Harlow: Longman.

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

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

Linacre, J.M. (2011). *Winsteps Rasch Measurement* (Version 3.72.3). www.winsteps.com. Author.

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

Meara, P.(1996).*The dimensions of Lexical Competence.* In G. Brown, K. Malmkjaer and J. Williams (eds.) *Performance and Competence in SecondLanguage Acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., & Jones, G. (1990).*Eurocentres Vocabulary Size Test, Version E1.1/K10.* Zurich: Eurocentres Learning Service.

Meara, P.,& Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142-151.

Messick, S.A.(1989). Validity. In Linn, R.L., editor, *Educational measurement*.3rd ed. New York: American Council on Education/MacmillanPublishing Company, 13–103.

Messick, S.A. (1995). Validity of psychological assessment: validation of inferences from persons'responses and performances as scientific inquiry into score meaning. *American Psychologist50*: 741-49.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York, NY: Newbury House.

Nation, I.S.P., (2001). *Learning Vocabulary in Another Language.* Cambridge University Press, Cambridge.

Nation I.S.P.,& Beglar, D. (2007). A vocabulary size test. *The Language Teacher 31*(7): 9-13.

Nation I.S.P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. London: Routledge.

Nation, P.(1990). *Learning and Teaching vocabulary*, New York, Newbury Houuse.

Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics*(pp. 35-54). New York: Oxford University Press Inc.

Nguyen, L.T.C, Nation, I.S.P. (2011) A bilingual vocabulary size test of English for Vietnamese learners.*RELC Journal 42*(1): 86-99.

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

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

Rahimi, M. (1996). *The study of English Language Instruction at the Secondary Schools of the Isfahan Province.* Unpublished M.A. Thesis, Shiraz University, Shiraz.

Razmjoo, S. A. & Riazi, M. (2006). Is communicative language teaching practical in the expanding circle? A case study of teachers of Shiraz high schools and institutes.*Journal of Language and Learning, 4,* 144-171.

Read, J. (2000). Assessing Vocabulary. Cambridge University Press, Cambridge, UK.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly, 18* (2), 199-219.

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

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329- 363.

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

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

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

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