Thi Qar Arts Journal
Vol 7 No 46 June. 2024

The Validation of The Vocabulary Level Test with The Rasch Model

Ahmed Sattar
stara6427@gmail.com
https://orcid.org/0009-0002-5684-3501

Omar Ahmed
omarahmednew57@gmail.com
https://orcid.org/0009-0004-8875-3919

https://doi.org/10.32792/tqartj.v7i46.607

Received 27/4/2024, Accepted 28/5/2024 , Published 30/6/2024

Abstract

The most significant feature of any prosperous measuring implement is validity. Briefly, This work outlines the six sides of the Messickian validity framework. A brief introduction is presented late, to the Rasch model—a measurement model used fields that are related to the society—is given. An input revealed the model of Rasch may sort to establish various validity's features that are taken from Messick's perspective has been examined, also Rasch analysis is used as an instructive case to show the validity of a vocabulary level test. The findings indicate that a few items don't fit the Rasch model. An analysis revealed that several items had distracters that should be eliminated because they were not working as planned. Also, the study demonstrated that the test covered a broad spectrum of the ability scale and was on target. The sample's performance on two test subsets showed that participants' measures on the two subsets were the same, proving the instrument's unidimensionality.

Keywords: vocabulary level test(VLT), Rasch model, Validity

Introduction

Vocabulary knowledge is important in future possibilities and people’s lives (Beck, McKeown, & Kucan, 2002). Following the performance of English second/foreign language readers’ encounter with different vocabulary; researchers have commented on the prominent role of vocabulary as an indicator of general reading skill (Nation, 2001). Indeed, ESL/EFL readers frequently stated lack of sufficient word understanding as one of the major barriers to content comprehension so vocabulary load is a very important cue of text complexity. Many vocabulary tests have been introduced by various writers .

The Vocabulary Levels Test (VLT) is perhaps One of the most frequent models of L2 that has been used to check vocabulary knowledge is possibly vocabulary levels test or VLT (Read, 2000). Nation (1983) was the first one to establish the model and later it was updated by Schmitt et.al (2001) to have the final say to the level to where examinees could the links that might extent to four meanings of words and range from two thousands to ten thousands number of words and an educational word level. The test can be done as a whole with Participants in the test can complete all stages as one way of doing the test, or students can individually participate in the test. Possibly, only necessary to It is only feasible to test the two thousands of beginners stage because participant probably won't master the test without proper instructions

The more frequent vocabulary that students should concentrate on means the higher value of VLT. Thirty questions are given to each level of the participants showing that the VLT uses a single shape format. The words are presented in ten clusters of 6 vocabularies , three of these are key vocabularies and three to distract the participants, and three meanings at every stage. Writing the correct numbers of items with its consistent descriptions is what the participants are required to do . The test takers gets a mark for every correct answer that they choose, that is the highest mark that they can get is 30. When scoring the test, the scores for the individual levels are most important because these scores reveal where subsequent vocabulary learning should be focused. In contrast, the overall score has little meaning. The items in 5 of the 10 clusters are made up of nouns. The items in 3 of the clusters are verbs, and the items in 2 of the clusters are adjectives. The proportion of nouns, verbs, and adjectives is representative of their proportional occurrence in English although it should be noted that this may vary within frequency bands.

Schmidt, Schmitt, & Clapham's (2001) new forms of the test improved on the earlier ones by increasing the number of items per level from 18 to 30 to improve reliability, and selecting academic words from Coxhead's (2000) Academic Word List rather than the source: Xue & Nation's (1984) University Word List. While these changes greatly improved upon the original version, Schmitt, Schmitt, & Clapham's VLT still had two limitations (Webb & Sasao, 2013). First, items within the word frequency levels were derived from texts from the 1930s and 1940s, and therefore might not reflect current vocabulary. Second, the earlier forms of the VLT did not measure knowledge of the most frequent 1000-word families. This is particularly important because the relative value of words has a marked decrease after the most frequent 1000-word families; the most frequent 1000-word families account for as much as 80% of English, while the most frequent 1001 to 2000-word families make up from around 4 to 10% of English. Thus, the most valuable word frequency level to measure is the most frequent 1000-word families because of its importance to understanding English.

Tests Validity

Various specialists at various steps of stages of time examined The validity. A well- established test is the one that measures the things that showed be measured (Kelly, 1927). After that, The American Psychological Association recognized the following kinds of validity: analytical, content , parallel, and construct validity. The one that is concerned with how many items that are used in the test is called Content validity . These are chosen from a group of vocabularies to what extent they can represent the content that should be tested. The one that is concerned with how effective a test is to anticipate the participants future performances and is measured by compare the scores that the testers gets with their scores that they might get in the future is called Predictive validity. When the correlation is high between the present and the future , the prediction validity will be the most reliable to depend on.

This combination was because both predictive and concurrent validity are computed by correlating the test in focus with another test set as a criterion. Thus, four types of validity were reduced to three main types: content, criterion-related, and construct validity. Gradually, theorists began to move in the direction of unifying the three types of validity into one type which was construct validity. For example, Cronbach (1980) mentioned that "all validation is one" (p. 99), and by "one" he meant construct validity. Finally, Messick (1989) confirming the unitary nature of validity, extended the definition of construct validity and defined it as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 288). For Messick (1989,1995), validity is a unitary concept realized in construct validity and has six facets content, substantive, structural, generalizability, external, and consequential.

The substantive aspect of construct validity may be roughly defined as the substantiation of the content aspect. It deals with finding empirical evidence to ensure that test-takers are engaged with the domain processes provided by the test items or tasks. An obvious example is multiple choice distracter analysis which is carried out to provide empirical evidence for "the degree to which the responses to the distracters are consistent with the intended cognitive processes around which the distracters were developed" (Wolfe & Smith, 2007, p. 209).

The consequential aspect of validity focuses on the value implications of score interpretation as a source for action. Evidence concerning the consequential aspect of validity also addresses the actual and potential consequences of test score use, especially regarding sources of invalidity such as bias, fairness, and distributive justice. (p.244) A simple example in which case consequential aspect of validity is violated could be a test that includes items that are biased in favor of a group of test takers and thus results in high scores for one group and low scores for the other.

RASCH MODEL

Attempts have been made to extend the current view of construct validity along with its six facets to the Rasch model framework. Bond (2003), Smith (2001), and Wolfe & Smith (2007) have all attempted to point out how the analyses carried out within the Rasch framework can be linked to current validity arguments. Rasch model, named after the Danish mathematician and statistician Georg Rasch, is a prescriptive probabilistic mathematical ideal. It is highly distinguished for its two remarkable properties of invariance and interval scaling which are obtained in case the basic assumption of unidimensionality underlying the model is met, i.e. when the data fit the model. The model is referred to as a prescriptive model because it prescribes specific conditions for the data to meet. This means that the whole research process, from the very beginning, must be in line with the model's specifications. One of the basic assumptions of the Rasch model is the unidimensionality principle: the measurement instrument must measure only one trait at a time.

Though theoretically sound, practically it is almost impossible to construct a test that measures only one attribute or to prevent the interference of extraneous factors. One may unintentionally measure language proficiency in a math test which is primarily intended to measure the test takers' mathematical ability. This is usually the case with math tests including worded problems, especially when the test is administered to non-native speakers of the test language. Moreover, in almost all testing situations, a number of extraneous factors are involved which contaminate the measurement. Henning et al. (1985) clarify the point: Examinee performance is confounded with many cognitive and affective test factors such as test wiseness, cognitive style, test-taking strategy, fatigue, motivation, and anxiety. Thus, no test can strictly be said to measure one and only one trait. (p. 142) As achieving this strong version of unidimensionality is impossible, a more relaxed formulation has also been advanced (Bejar, 1983). The unidimentionality with which the Rasch and IRT models are concerned is psychometric unidimensionality and not psychological. Thus unidimentionality within Rasch model 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. 27).

As achieving this strong version of unidimensionality is impossible, a more relaxed formulation has also been advanced (Bejar, 1983). The unidimentionality with which the Rasch and IRT models are concerned is psychometric unidimensionality and not psychological. Thus unidimentionality within Rasch model 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 order for the data to meet unidimensionality condition, the response patterns should follow Guttman pattern. If items are ranked from easy to difficult, a person who has responded correctly to an item should reply correctly to all the easier items as well. In other words, it is not expected that a person respond correctly to difficult items, but miss the easier ones or vice versa. The more the data is Guttman-like, the more it is likely to fit the Rasch model. Having calculated the probabilities of providing correct responses to items of specific estimated difficulties by persons of particular estimated abilities, one should check whether the model's expectations realized in the form of probabilities are consistent enough with the observed data. This is done by checking the probabilities against the real observed data which can be carried out statistically as well as graphically. It should be noted that there always exists some difference between the model's predictions and the real data since the model is a perfect mathematical ideal, a condition impossible to meet in the real world.

If deviation of data from the ideal set by the model is tolerable, it is said that the data fit the model, thus enabling one to benefit from the attractive properties provided by the model. If not, the remarkable properties of the model which are in fact the properties of fundamental measurement are lost. Although over forty fit indices have been developed by Psychometricians to check the accord between data and the model mainly two of them are implemented in Rasch software written in North America and Australia: infit and outfit statistics. While the former is sensitive to unexpected patterns of response in the zones where the items are quite targeted to the person's abilities, the latter is highly sensitive to lucky guesses and careless mistakes. Both types of fit statistics are expressed in the form of mean square values as well as standardized values. The ideal value is 1 for mean square values and 0 for standardized ones. The acceptable range for mean square values is from 0.70 to 1.3 and for standardized ones from -2 to +2. In case the data fit the model, one can be confident that the item measures are independent of the person measures and vice versa.

The invariance of the measures can also be tested by splitting the items or persons into two halves and running independent analyses to check whether the item and person estimates remain invariant across the analyses. To be more specific, either the same test is given to two groups of people, or the sample to which the test is given is divided and considered as two groups. Then, the difficulty estimates of each item, derived from two separate analyses, are plotted against each other on x and y axes. The procedure is the same for persons, but in this case of persons, there are two groups of items and one group of persons. That is, two ability estimates for each person are estimated based on the two sets of items and then the ability estimates are plotted against each other. A dotted line which indicates "the modeled relationship required for invariance" (Bond & Fox, 2007, p. 72) is drawn and 95% of control lines based on standard errors of item or person pairs are constructed around it. The items or persons falling between the control lines are considered to be invariant.

Regarding Rasch analysis and validity, the works of Bond (2003), Smith (2001), and Wolfe and Smith (2007), the contribution that Rasch analysis can make to demonstrate different aspects of construct validity is pointed out. Several analyses are performed to provide evidence for the content aspect of validity within the Rasch framework. Fit indices are used to check the relevance of the test content to the intended construct. Misfitting items may be measuring a totally different and irrelevant construct. Moreover, person-item maps and item strata are two important criteria for checking the representativeness of the items. Noticeable gaps in the item difficulty hierarchy point to the fact that some areas of the construct domain have not been covered by the test (Baghaei, 2008). Item strata, i.e. "the number of statistically distinct regions of item difficulty that the persons have distinguished" (Smith, 2001, p. 293), is another clue that is drawn upon to check representativeness. There should be at least two item difficulty levels distinguished to judge the items as being appropriate representatives of the intended content.

Furthermore, technical quality of the test items can be assessed via fit indices as well as item-measure correlations since the former is a good indicator of multidimensionality, poor item quality or miskeying and the latter is an indicator of "the degree to which the scores on a particular item are consistent with the average score across the remaining items." (Wolfe & Smith, 2007, p. 206). With regard to the expected values of the item-measure correlations, Wolfe and Smith (2007) summarize the issue as: Item-measure correlations should be positive, indicating that the scores on the item are positively correlated with the average score on the remaining items. Negative item-measure correlations typically indicate negatively polarized items that were not reverse- scored. Near zero item-measure correlations typically indicate that the item is either extremely easy or difficult to answer correctly or to endorse or that the item may not measure the construct in the same manner as the remaining items. (p. 206)

METHOD

Regarding the Participants, thirty undergraduate students at Thi- Qar University were randomly selected. Their age ranges from 19 to 25. Gender and language background were not used in the selection procedure.

For Instruments, 30 (VLT) of English vocabulary was given to the participants. Time allowed for answering all the items was 45 minutes though some of the participants finished the test sooner.

RESULTS

The data were analyzed using WINSTEPS Rasch software version 3.66.0 (Linacre, 2008). First of all, fit indices were examined closely to check the relevance of the items as part of content validity. Table 1 shows the fit indices for some of the items. The items are arranged from difficult to easy. The first column, "ENTRY Number", indicates the number given to each item in the test (ranging from 1 to 30). 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). The number of participants who have attempted each item is given in the third column which is labeled as "COUNT". The difficulty estimates for the items are given in the fourth column labeled as "MEASURE".

Table1. Item Statistics: Measure Order

	Total Score	Count	Measure
Mean	27.3	30.0	.0
S.D	12.3	.0	1.21

Table 1 indicates many Items should be either omitted or revised because of lack of fit to the model. These items are measuring something other than the intended content and construct. That is, they are constructirrelevant.

Having a look at table 2, the Summary Statistics, one can investigate the representativeness of the items by checking the value given for item strata. Item strata is labeled as "SEPERATION" in the table. The minimum value for item strata is 2. The separation value given for this test is 3.38 which is an acceptable index. Thus, one can rely on the representativeness of the test items.

Table2. SUMMARY OF 30 MEASURED ITEMS

	Total score	Count	Measure	Model error
Mean	27.3	30.0	.00	.32
S.D	12.3	.0	1.21	.11
Max.	29.0	30.0	2.27	1.01
Min.	4.0	30.0	-4.52	.28

Looking at the Previous tables, the test is just fairly good as far as external aspect of validity is concerned. The test is very well-targeted for the sample. Had we given this test to an untreated group, it would not have been capable of detecting changes in the high-ability persons after the treatment as the dispersion of item calibrations beyond the highest person measure is not very wide. More items would have been needed to cover the area beyond the highest person measure. Having a look at the moderate person strata value (2.50) confirms this point. However, since the test is not constructed for purposes of detecting changes after treatment, this lack of floor effect does not pose a problem.

To check the invariance of person measures and provide evidence for generalizability aspect of validity, the items are divided into two halves. Then for each person, two ability measures are estimated and plotted against each other (Baghaei, 2009). The manifested invariance of person measures provides evidence for the generalizability aspect of construct validity.

CONCLUSION

In this paper an overview of validity from Messick's viewpoint was provided. Afterwards, the Rasch model as a new measurement theory was introduced. Rasch model, rejecting the concept of raw score, provides person and item estimates that are placed on an interval scale and thus is a more appropriate model than the classical test theory for measurement in the human sciences.

It was then indicated that it is possible to extend the Messickian view of validity to the Rasch model. Various analyses within Rasch model were mapped to

different aspects of validity. The possibility of demonstrating the validity of measuring instruments makes the Rasch model a valuable tool for construct validation of tests. It was shown that it is possible to link the Messickian view of construct validity with its six facets (content, substantive, structural, generalizability, external and consequential) as defined by Messick to several analyses available within Rasch model framework. A vocabulary levely test was used to empirically apply the Rasch model analyses for validation. The results show that Rasch model works well for establishing the validity of language tests and can routinely be used by language testing specialists to provide validity evidence for their tests.

References

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

Beck, I.L., McKeown, M.G., & Kucan, L. (2002). Bringing words to life: Robust vocabulary instruction. New York: Guildford Press.

Bejar, I.I. (1983). Achievement testing: recent advances. Beverly Hills, CA Sage.

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

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

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213–238.

Cronbach, L. J. (1980). Validity on parole: how can we go straight? New directions for testing and measurement: measuring achievement over a decade. Proceedings of the 1979 ETS Invitational Conference. San Francisco: Jossey-Bass.

Henning, G., Hudson, T. & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language test. Language testing, 2:2, 141-154.

Kelly, T. L. (1927). Interpretation of educational measurements. New York: Macmillan.

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

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

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

Messick, S. (1996). Validity and washback in language testing, Language Testing, 13:3, 241-256.

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

Read, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University Nation, I. S. P. (1983). Testing and teaching vocabulary. Guidelines, 5(1), 12–25.Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language Testing, 18(1), 55–88.

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

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

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. RELC Journal, 44(3), 263– 278.

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.

Xue, G., & Nation, I. S. P. (1984). A university word list. Language Learning and Communication, 3(2), 215 229.