ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 2, No. 5, pp. 1052-1060, September 2011

doi:10.4304/jltr.2.5.1052-1060

Validation of a Multiple Choice English Vocabulary Test with the Rasch Model

Purya Baghaei
Islamic Azad University, Mashhad Branch, Iran
Email: puryabaghaei@gmail.com

Nazila Amrahi
Urmia University, Iran

***Abstract*—Validity is the most salient characteristic of any effective measurement instrument. This paper briefly reviews the Messickian validity framework along with its six facets. Then the Rasch model as a measurement model which yields interval scaling in the social sciences is briefly introduced. The contribution that Rasch model can make to establishing different aspects of validity from Messick's view point is discussed and as an illustrative example, the validity of a multiple-choice test of vocabulary is demonstrated using Rasch analysis. The results show that several items misfit the Rasch model. Distractor analysis showed that a few items have distracters which do not function in the intended way and need to be removed. Item-person map showed that the test was on-target and covered a wide range of the ability scale. The performance of the sample on two subsets of the test revealed that participants had identical measures on the two subsets, which is evidence of the unidimensionality of the instrument.**

***Index Terms*—validity, Rasch model, fit statistics, item-person map, unidimensionality**

I. Validity

Validity of a test has been defined in a number of ways by different scholars at different stages of time. For Kelly (1927), a valid test is a test which measures what it is intended to measure. Later on, in 1954, the American Psychological Association distinguished four types of validity: content, predictive, concurrent, and construct validity. Content validity is concerned with the extent to which the items included in a test are selected from a universe of items and the extent to which they are representative of the content intended to be tested. Predictive validity is considered as the effectiveness of a test to predict the test takers' future performance and is calculated via correlating the results of the intended test with another test given in some future time; the higher the correlation, the greater the predictive validity of the test would be. Concurrent validity is very similar to predictive validity in that it is concerned with the degree of correlation with another test, the difference being that the criterion test is given at approximately the same time. Concurrent validity is required to substitute a test for an already existing standard one due to practicality issues. Finally, construct validity is concerned with the extent to which a test is reflective of the underlying construct the test is supposed to assess. Later, predictive and concurrent validity were combined into one type of validity, namely criterion-related validity (Smith, 2001). This combination was due to the fact that 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 the 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 on the basis of 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 of content, substantive, structural, generalizability, external, and consequential.

The content aspect of construct validity mainly refers to content relevance, representativeness, and technical quality. The concern of content validity is that all the items or tasks as well as the cognitive processes involved in responding to them be relevant and representative of the construct domain to be assessed. The extent to which test items or tasks are relevant and representative of the construct domain is normally determined by professional judgment of experts. Technical quality of items, referring to issues like "appropriate reading level, unambiguous phrasing and correct keying", (Messick, 1996, p. 248) is also considered to be part of the content aspect of construct validity.

Substantive aspect of construct validity may be roughly defined as the substantiation of the content aspect. It deals with finding empirical evidence to assure that test-takers are actually 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).

Structural aspect of construct validity is mainly concerned with the scoring profile. It is highly important to take into account the structure of the test when scoring it. It does not seem sound to add up the scores of the different parts of a test, when each part measures a different dimension. While one single score can summarize the performance of an individual on a unidimensional test, scores on different dimensions must be reported separately. In other words, the scoring models should be informed by the structure of the test.

Generalizability aspect of construct validity deals with the extent to which the score meanings and interpretations are generalizable to other tasks and contents which are not included in the test but are part of the broader construct domain. In other words, the generalizability aspect tells us to what extent we can depend on the test scores as broader indicators of a person's ability and not just as an index of the examinee's ability to perform a limited number of tasks included in an assessment device.

The external aspect of construct validity is concerned with the degree to which test scores are related to other test and non-test behaviors. In Messick's (1996) own words:

The *external* aspect of validity refers to the extent to which the assessment scores' relationships with other measures and non-assessment behaviors reflect the expected high, low and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations. (p. 251)

Analyses of group differences and responsiveness of scores to experimental treatment (Messick, 1989) are considered to be two important methods which serve as important evidence for the external aspect of construct validity. If a measurement instrument shows sensitivity to changes in the test takers' levels of latent trait as a result of introducing treatment as an external non-assessment behavior, it is said to have external validity. That is, if a test is given before and after a treatment and results indicate that the test-takers did better on the test after the treatment, it is said the test has external validity. Moreover, a test is said to have external validity in case it can differentiate between those who possess the construct and those who do not or between those who possess varying levels of the construct.

The consequential aspect of construct validity, as the name suggests, deals with the intended and unintended (e.g. bias) consequences of the assessment and the implications scores meanings have for action. According to Wolfe and Smith (2007):

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 in regard to 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 which 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.

II. Rasch Model Introduction

Attempts have been made to extend the current view of construct validity along with its six facets to Rasch model framework. Bond (2003), Smith (2001), and Wolfe & Smith (2007) have all attempted to point out how the analyses carried out within 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 which 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) clarifies 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. 271).

In order for the data to meet unidimensionality condition, the response patterns should follow Guttman pattern. If items are rank ordered 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 the 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.

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 is 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% 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.

III. Rasch Analysis and Validity

In this part, summarizing briefly 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.

A number of analyses are performed to provide evidence for the content aspect of validity within 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 map 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 area of the construct domain has 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 which is drawn upon to check representativeness. There should be at least two item difficulty levels distinguished so as 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)

Person fit statistics and, in the case of multiple-choice tests, multiple choice distracter analysis are considered to be important indicators of substantive aspect of validity. Person fit statistics provide empirical clues for "the extent to which a person's pattern of responses to the items correspond to that predicted by the model" (Smith, 2001, p. 296). Person misfit may be due to factors like carelessness, guessing, etc. Distracter analysis within Rasch framework involves distracter p-valuse, choice means and distracter-measure correlations. P-values indicate "the proportion of respondents choosing each distracter" (Wolfe & Smith, 2007, p. 209). Ideally, it is expected that the distracters be equally attractive; however, this seems to be almost impossible in practice. Thus, p-values are used to detect malfunctioning as well as non-functioning distracters. Choice means represent "the average measure of respondents who choose each distracter" (Wolfe and Smith, 2007, p. 209). They indicate the discrimination power of the distracters.

It is expected that distracters be chosen by less able test takers, thus discriminating between test takers of high and low ability levels. As Wolfe and Smith (2007) put it," If a distracter does not attract less able respondents, then its validity as a measure of the underlying construct is questionable." (p. 209). Finally, distracter-measure correlations are correlations between distracters and test takers' ability measures and indicate "the degree to which each distracter is a plausible answer to the prompt" (Wolfe and Smith, 2007, p. 209). Since, again, it is expected that test takers of low ability choose the distracters (rather than a correct option), thus negative values for correlations are desired. However, since the number of test takers choosing a particular distracter may be small, it is likely that the distracter measure correlations be attenuated and consequently result in correlation values which are not considerably negative. In such cases, choice means are drawn upon to compensate for the attenuation effect.

Fit statistics are used to assure whether the test is unidimensional and guide one to decide upon the way the test should be scored. That is, in case the test is shown to be unidimensional, reporting a single score for the whole test would suffice. However, in case of multidimensionality, separate scores should be reported for each dimension, and one should be cautious not to add up the scores on different dimensions. Thus, fit statistics provide helpful evidence with regard to the structural aspect of construct validity.

Checking the invariance of item measures across different populations or over time, as well as checking the invariance of person measures across different sets of items can be employed to check the generalizability aspect of construct validity.

In the case of external aspect of construct validity, the extent to which the meanings of the scores of a test hold relations with some other related test results or non-test behaviors is usually checked via building Multitrait-Multimethod matrices. The external aspect of validity is usually checked via monotrait and heterotrait correlations which have traditionally been referred to as convergent and discriminant evidence respectively. It is expected that monotrait correlations be higher than the heterotrait ones in order to serve as evidence for the external aspect of validity. Moreover, the capacity of a test to detect within-individual changes (over time, e.g. as a result of treatment) and between-group differences, is another indicator of the external validity. This capacity can be checked via visual inspection of person-item map as well as checking the person strata index. If a test is given to a group before a treatment and the map manifests "a floor effect, and a wide dispersion of item calibrations beyond the highest person measure" (Wolfe & Smith, 2007, p. 222), the test is said to be responsive to treatment and thus capable of detecting within-individual changes. The same applies to situations where the test is used to compare different groups which undergo different experimental treatments. Person strata index which represents the number of statistically separate ability strata that the test can distinguish is another evidence for external aspect of construct validity. High values for person strata (at least 2) are needed to confirm the external aspect of validity of a test.

Rasch has not explicitly put forward a way to check the consequential aspect of validity. However, issues like item bias and examination of differential item functioning (DIF) or a close examination of the person-item map- which reveals the amount of information on the basis of which decisions for action are taken- can provide helpful evidence to decide about the consequential aspect of construct validity of a test.

In the following section the analyses discussed above are applied to a multiple-choice English vocabulary test to demonstrate its construct validity. Rasch analyses corresponding to various Messickian validity aspects are conducted to show how Rasch model is applied in practice for validation.

IV. Method

A. Participants

Sixty undergraduate English Language and Literature students at Urmia University were randomly selected. Their age ranges from 19 to 25. Gender and language background were not used in the selection procedure.

B. Instruments

A 70-item multiple choice test of English vocabulary was given to the participants. They were required to choose the best possible answer for each item. Time allowed for answering all the items was 45 minutes though some of the participants finished the test sooner.

C. 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 70). 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". The fifth column, "MODEL S. E.", shows 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 "OUTFIT" as well as "INFIT" columns.

TABLE 1
ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	ITEM	G
42	4	60	2.72	.53	1.09	.3	1.40	.8	.04	.20	Item 42	0
19	11	60	1.50	.35	1.20	1.0	1.26	.9	.05	.29	Item 19	0
22	13	60	1.26	.33	1.12	.7	1.27	1.1	.14	.31	Item 22	0
57	14	60	1.16	.32	1.03	.2	1.19	.8	.25	.32	Link 7	0
25	16	60	.96	.31	1.14	.9	1.28	1.3	.13	.33	Item 25	0
48	16	60	.96	.31	1.16	1.1	1.16	.8	.15	.33	Item 48	0
56	18	60	.77	.30	1.29	2.0	1.33	1.7	-.01	.34	Link 6	0
34	19	60	.69	.30	1.10	.8	1.11	.7	.22	.34	Item 34	0
69	22	60	.43	.29	1.11	1.0	1.17	1.2	.21	.35	Link 19	0
9	23	60	.35	.28	.93	-.6	.93	-.5	.43	.35	Item 9	0
61	23	60	.35	.28	1.35	3.0	1.40	2.8	-.07	.35	Link 11	0
14	24	60	.27	.28	1.01	.1	1.00	.0	.34	.35	Item 14	0
40	24	60	.27	.28	1.07	.7	1.12	.9	.26	.35	Item 40	0
18	25	60	.19	.28	1.10	1.0	1.10	.8	.24	.35	Item 18	0
31	25	60	.19	.28	1.15	1.5	1.20	1.6	.16	.35	Item 31	0
10	26	60	.12	.28	.79	-2.3	.76	-2.3	.61	.35	Item 10	0
28	26	60	.12	.28	.92	-.8	.90	-.9	.45	.35	Item 28	0
62	30	60	-.19	.28	.94	-.7	.92	-.7	.43	.35	Link 12	0
64	30	60	-.19	.28	.73	-3.2	.70	-3.1	.67	.35	Link 14	0
3	31	60	-.27	.28	1.00	.0	1.02	.2	.34	.35	Item 3	0
37	31	60	-.27	.28	.89	-1.3	.88	-1.1	.48	.35	Item 37	0
67	35	60	-.57	.28	1.13	1.4	1.11	.9	.19	.34	Link 17	0
1	38	60	-.81	.28	1.11	1.1	1.15	1.0	.19	.33	Item 1	0
47	38	60	-.81	.28	1.00	.0	1.04	.3	.32	.33	Item 47	0
5	39	60	-.89	.29	.85	-1.4	.78	-1.5	.52	.33	Item 5	0
66	39	60	-.89	.29	.76	-2.4	.68	-2.3	.63	.33	Link 16	0
36	40	60	-.97	.29	.79	-2.0	.72	-1.9	.58	.32	Item 36	0
63	40	60	-.97	.29	.91	-.8	.88	-.7	.44	.32	Link 13	0
65	44	60	-1.33	.31	.99	.0	.98	.0	.31	.30	Link 15	0
54	54	60	-2.60	.44	1.08	.3	1.35	.8	.05	.20	Link 4	0
51	58	60	-3.80	.72	.96	.2	.52	-.3	.23	.12	Link 1	0
52	59	60	-4.52	1.01	.98	.3	.53	.0	.15	.08	Link 2	0
MEAN	27.3	60.0	.00	.32	1.00	-.1	1.00	.0				
S.D.	12.3	.0	1.21	.11	.12	1.1	.20	1.1				

Acceptable values range from 0.7 to 1.3 for "MNSQ" and from -2 to +2 for "ZSTD". The "PT-MEASURE" column indicates the observed ("CORR.) as well as the expected ("EXP.) correlation between performance on each item and the ability estimates of the participants who have responded correctly to that item. Finally, the last column shows the labels given to the items by the analyst.

Table 1 shows that Item 42 is the most difficult item on the test. From 60 participants who have attempted this item, only 4 could get it right. The difficulty of this item is estimated to be 2.72 logits with the standard error of 0.53. This means one can be 95% sure that the true value for the difficulty of this item lies somewhere between 1.66 to 3.78 logits, i.e., two SE's below and above the observed measure. The infit indices are within the acceptable range. Though the MNSQ for outfit has exceeded the limit a little bit, the ZSTD is within the acceptable range, thus not causing a serious problem. There is some difference between the observed and expected correlation between the performance on this item and the participants' ability measures on the entire test which can be associated with the observed deviation in outfit MNSQ.

Table 1 indicates that items 56, 61, 10, 64, 66, and 36 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 construct-irrelevant.

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.

TABLE 2
SUMMARY OF 70 MEASURED ITEMS

	RAW SCORE	COUNT	MODEL MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	27.3	60.0	.00	.32	1.00	-.1	1.00	.0
S.D.	12.3	.0	1.21	.11	.12	1.1	.20	1.1
MAX.	59.0	60.0	2.72	1.01	1.35	3.0	1.40	2.8
MIN.	4.0	60.0	-4.52	.28	.73	-3.2	.52	-3.1
REAL RMSE	.34	ADJ.SD	1.16	SEPARATION	3.38	ITEM	RELIABILITY	.92
MODEL RMSE	.34	ADJ.SD	1.17	SEPARATION	3.45	ITEM	RELIABILITY	.92
S.E. OF ITEM MEAN = .15								

Moreover, person-item map (Figure 1) can serve to provide evidence for the representativeness of the test items, that is, content validity. The map shows that the bulk of items on the right are matched to the bulk of persons on the left, indicating the test is appropriately targeted for this group of participants. Link 1, Link 2, and Link 4 (i.e. items 51, 52, and 54), though indicating good fit to the model, can be omitted since they are too easy for the participants and are in fact useless because there are not any participants at that ability level. If we had some test takers at the lower end of the scale we would need to add some more items at that difficulty level to cover the gap between items labeled Link1 and Link 4 to measure the ability of the persons in that region of the scale more precisely. There is also a gap between items 44 and 17, meaning some items are needed to cover this area of the construct domain. Omitting Link 1, Link 2, and Link 4, the rest of the items can be said to be on target. Moreover, there is no ceiling effect. There are items whose difficulty levels are above the most able participants' ability levels. Overall, the items show acceptable degree of representativeness.

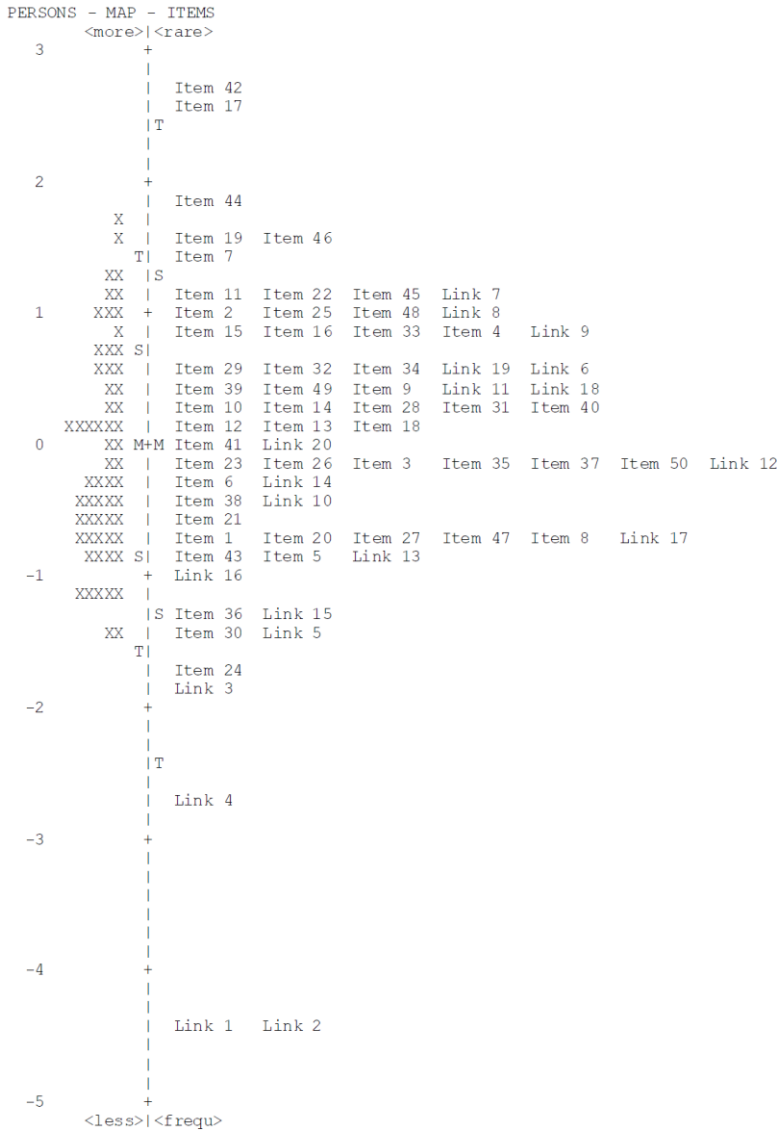

Figure 1 – PERSON-ITEM MAP

Table 3, the "Item Distracter Table" reveals helpful information regarding substantive aspect of construct validity. Item Distracter Table for 10 items is given below. The first column shows the entry number for each item. The second column, "DATA CODE", indicates the codes given to the options of the item. That is, 1 represents the first choice, 2 the second, etc. "." represents the missing values, i.e. the cases wherein none of the options was chosen. "SCORE VALUE" column shows the correct option by coding it as 1 and the other incorrect options as 0. The fourth column, "DATA COUNT", indicates the number as well as the percentage of the participants who have chosen a particular option, be it right or wrong. "AVERAGE MEASURE" or choice mean shows the mean of the ability estimates of all the participants who have chosen a particular option. It is expected that the value for average measure be the highest for the correct option and lower for incorrect options. There is an asterisk placed above the average measure for correct options in cases where this expectation is not met. This is the case with items 19, 22 and 57 as shown in Table 3. "S. E. MEAN", is

an abbreviation for standard error of the mean which is the standard error of the mean of the ability estimates of those participants who have chosen a particular option. Finally, "PTMEA CORR." shows the correlation between the occurrence and non-occurrence of each option and the ability estimates of the participants choosing a particular option.

TABLE 3
ITEM DISTRACTER

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	DATA %	AVERAGE MEASURE	S.E. MEAN	PTMEA CORR.	ITEM
42	3	0	13	22	-.29	.25	-.08	Item 42
	1	0	8	13	-.23	.19	-.03	
	.	0	4	7	-.21	.17	-.01	
	2	0	11	18	-.18	.20	.00	
	5	0	20	33	-.10	.22	.07	
	4	1	4	7	-.06	.39	04	
17	.	0	3	5	-.51	.54	-.10	Item 17
	3	0	19	32	-.40	.18	-.19	
	5	0	5	8	-.39	.29	-.08	
	2	0	25	42	-.16	.15	.02	
	1	0	3	5	.32	.10	.15	
	4	1	5	8	.68	.34	.33	
44	3	0	11	18	-.67	.16	-.29	Item 44
	.	0	2	3	-.49	.97	-.07	
	5	0	23	38	-.27	.16	-.09	
	1	0	6	10	-.18	.36	.00	
	4	0	9	15	-.06	.15	.06	
	2	1	9	15	.59	.27	.41	
19	1	0	1	2	-1.36		.20	Item 19
	2	0	21	35	-.36	.17	-.17	
	.	0	3	5	-.09	.36	.03	
	5	0	22	37	-.06	.20	.12	
	3	0	2	3	.31	.47	.12	
	4	1	11	18	-.10*	.18	.05	
46	.	0	1	2	-1.46		-.21	Item 46
	2	0	17	28	-.50	.15	-.26	
	4	0	19	32	-.29	.15	-.09	
	1	0	4	7	-.16	.32	.01	
	3	0	8	13	-.08	.31	.05	
	5	1	11	18	.53	.26	.43	
7	2	0	9	15	-.93	.18	-.40	Item 7
	.	0	1	2	-.75		-.09	
	4	0	3	5	-.71	.21	-.16	
	1	0	10	17	-.24	.20	-.04	
	3	0	24	40	-.08	.16	.10	
	5	1	13	22	.37	.19	.37	
22	5	0	9	15	-.50	.17	-.17	Item 22
	.	0	9	15	-.48	.22	-.16	
	2	0	4	7	-.41	.26	-.08	
	4	0	15	25	-.35	.22	-.13	
	1	0	10	17	.45	.23	.36	
	3	1	13	22	.03*	.23	.14	
57	.	0	5	8	-.79	.41	-.24	Link 7
	1	0	20	33	-.34	.14	-.14	
	4	0	20	33	-.15	.17	.03	
	2	0	1	2	.34		.09	
	3	1	14	23	.17*	.25	.25	
58	1	0	11	18	-.44	.22	-.16	Link 8
	4	0	19	32	-.35	.16	-.15	
	.	0	7	12	-.15	.24	.02	
	3	0	9	15	-.11	.27	.04	
	2	1	14	23	.19	.25	.26	
2	3	0	7	12	-.59	.17	-.19	Item 2
	4	0	7	12	-.27	.36	-.04	
	.	0	5	8	-.20	.27	-.01	
	5	0	6	10	-.18	.25	.00	
	2	0	20	33	-.18	.18	.00	
	1	1	15	25	.05	.24	.17	

Although all the information provided by the table is helpful, special attention is given to average measures. Items whose correct options are marked with asterisks (as a sign of flagging unacceptable values) should be checked. Those which have good fit indices and also the average measures for their wrong options are smaller than the average measure for their correct option are kept. However, those manifesting poor fit and those with greater average measures for wrong options than the correct option should be revised or deleted. Putting aside the items which do not fit the model, items 22, 19 and 57 have the asterisk above their correct options. This means that the mean of the persons who have chosen the right option is not greater than the means of those who have chosen the wrong options. This indicates that these distractors do not function in the expected fashion.

Distracter analysis showed that the distracters of most items acted in the intended way, i.e. elicited responses consistent with the intended cognitive processes, to a great degree. This is how multiple choice distracter analysis provides empirical evidence for the substantive aspect of construct validity.

Looking at the person-item map, 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). As is clear from Figure 2, all the persons are placed between the two control lines. The manifested invariance of person measures provides evidence for the generalizability aspect of construct validity.

Implications for the consequential aspect of construct validity can be drawn from the item-person map. Since there do not exist considerable gaps in the item hierarchy where the persons are located on the map, it seems that one can be somehow confident about the decisions made on the basis of this test. This is because the results are based on sufficient amount of information since the items are targeted to the ability levels of almost all the participants. This is relevant to the consequential aspect of validity.

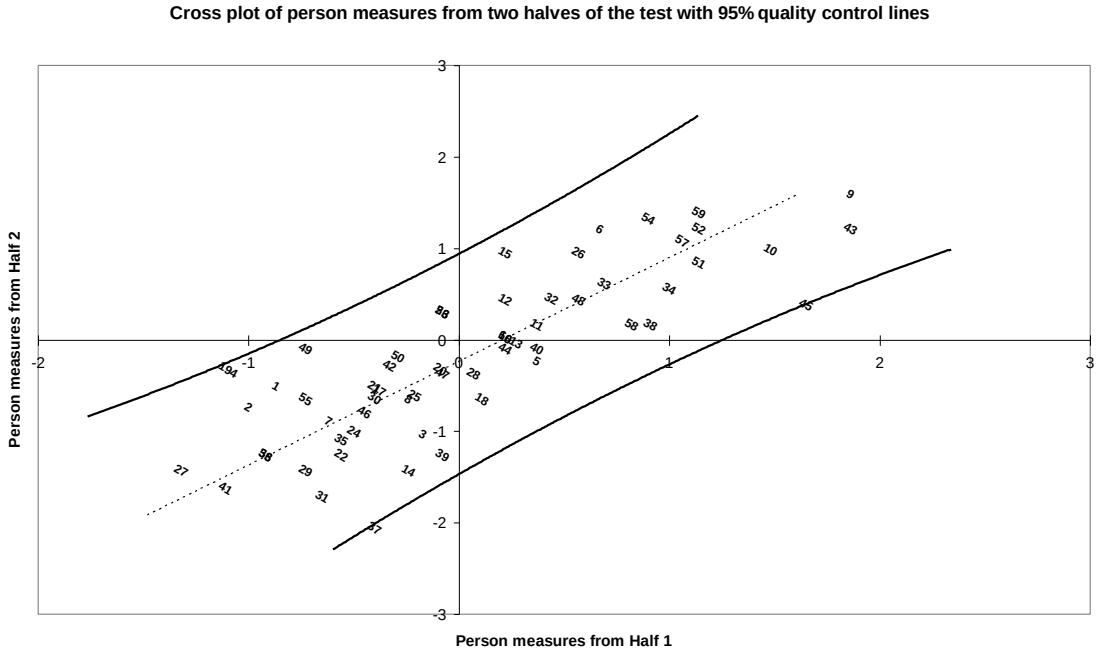

Figure 2 – Cross plot of person measures from two halves of the test with 95% quality control lines

V. 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 multiple-choice English vocabulary 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

[1] American Psychological Association. (1996). Standards for educational and psychological tests and manuals. Washington DC.
[2] Andrich, D. (1988). Rasch models for measurement. Newbury Park, CA: Sage.
[3] Baghaei, P. (2009). Understanding the Rasch model. Mashad: Mashad Islamic Azad University Press.

[4] 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.
[5] Bejar, I.I. (1983). Achievement testing: recent advances. Beverly Hills, CA Sage.
[6] Bond T. G. & Fox, C.M. (2007). (2nd ed.) Applying the Rasch model: fundamental measurement in the human sciences. Lawrence Erlbaum.
[7] Bond, T. G. (2003). Validity and assessment: a Rasch measurement perspective. *Metodologia de las Ciencias del Comportamiento* 5:2, 179-194.
[8] 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.
[9] Embretson, S. E., & Reise, S.P. (2000). Item response theory for psychologists. Hillsdale, NJ: Erlbaum.
[10] Gustafsson, JE. (1977). The Rasch model for dichotomous items: theory, applications and a computer program. Report No. 63. Institute of Education, University of Goteberg.
[11] Henning, G., Hudson, T. & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language test. *Language testing*, 2:2, 141-154.
[12] Kelly, T. L. (1927). Interpretation of educational measurements. New York: Macmillan.
[13] Linacre, J. M. (2004). Test validity and Rasch measurement: construct, content, etc. *Rasch Measurement Transactions*, 18:1, 970-971. Available: htpp://www.rasch.org
[14] Linacre, J. M. (2007). A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs. Chicago, IL: winsteps.com.
[15] McNamara, T. F. (1996). Measuring second language performance. New York: Longman.
[16] Messick, S. (1996). Validity and washback in language testing, *Language Testing*, 13:3, 241-256.
[17] Messick, S. (1989). Validity. In R.L. Linn (ed.) *Educational measurement* (pp. 13-103). New York: Macmillan.
[18] Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.
[19] 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.
[20] 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.
[21] Wright, B. D. & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Purya Baghaei is an assistant professor in the English Department of Mashhad Islamic Azad University in Iran. He holds a PhD in applied linguistics from Klagenfurt University, Austria. His major research interests are language testing and the application of Rasch models in education and psychology. He has published a book on fundamentals of Rasch model and has published several articles on language testing and Rasch measurement.

Nazila Amrahi holds an MA in English Language Teaching from Urmia University, Iran. Her main research interests are language testing and the applications of Rasch modeling in social science research.