International Journal of Language Testing
Vol. 6, No. 1, March 2016 ISSN 2476-5880

Examining Construct Validity of the Master's UEEUsing the Rasch Model and the Six Aspects of the Messick's Framework

Hamdollah Ravand[1], Tahereh Firoozi[2]

Received: 6 December 2015 Accepted: 13 February 2016

Abstract

The purpose of the present study was to explore validity of the University Entrance Examination for applicants into English Master's Programs in Iran (Master's UEE) with respect to the Messickian validity framework using the Rasch model. The data of 18821 test takers taking the 2009 version of the UEE were analyzed with Winsteps and R package eRm. An array of tests including Anderson's (1973) test were used to check unidimensionality of the test. Since the test as a whole did not show unidimensionality, the reading, grammar, and vocabulary sections of the test were analyzed separately through Anderson's (1973) likelihood ratio (LR) test using R package eRm. The results showed that (a) all the items in all the sections displayed good fit to the model, whereas more than 5 % of the examinees misfit the model, (b) due to small variance of item and person measures, the Rasch model explained small amount of variance in each section, namely 19.7 %, 13.8%, and 22.1 % in the reading, grammar, and vocabulary sections, respectively, (c) item measures were invariant within sections, contributing to the predictive validity (in the traditional sense of validity as types) of the test, whereas person measured did not show invariance, suggesting multidimensionality of the data hence threatening the construct validity of the test , (d) the bulk of the items did not match the bulk of the persons and there were noticeable gaps in the person-item maps, (e) small variance of person and item measures resulted in low Rasch reliability estimates for the sections, namely .53, .54, and .45 for the reading, grammar, and vocabulary sections, respectively.

Keywords: *Rasch model, Test validity, Unidimensionality, UEE*

1. Intoduction

Validity is a fundamental characteristic of any effective measurement instrument. As Bachman (1990) suggests, the primary concern in test development and test useis demonstrating that the interpretations and uses we make of test scores are valid. The concern with validity increases when a measure is used to make decisions that have serious consequences for stake holders. Such a concern has received due attention by policy-makers,administrators, and testing professionals

[1]*Vali-e-Asr University, Rafsanjan, Iran, University of Jiroft, Jiroft, Iran (Corresponding author, ravand@vru.ac.ir, ravand@ujiroft.ac.ir)*

2Vali-e-Asr University, Rafsanjan, Iran

in recent years so that different researchers have studied the validity of high-stake tests such as TOEFL (e.g., Stricker, L. J., & Rock, D. A., 2008; Bailey, K. M., 1999; Wall, D., &Horák, T., 2008) and SAT (e.g., College EntranceExamination Board, 2001; Kanarek, 1988; Bridgeman, &Wendler, 1991). Iranian University Entrance Exam for candidates into English master's programs (Master's UEE)is one ofthe most competitive tests in Iran.Each year more than one million applicantscompete to obtain a seat at one of thestate universities. The selection procedure at Iranian universities is carried out through university entrance examinations (UEE) administered for different levels of education namely bachelor, master's, and PhD. Besides being an academic competition, the UEE has turned into a seal of prestige in the Iranian society. UEE powerfully influences many spheres of social, academic, and educational lives of the applicants that the event could be a relevant subject to current developments and challenges in language testing and assessment. Consequently, the construct validity of this test needs due attention. Although some sporadic validityinvestigations (Barati&Ahmadi, 2010; Razavipour, 2010; Tahmasbi&Yamini, 2012) have been conducted on the UEE test at B.A. levels, to the best knowledge of the authors, no single study about the validity of the present test has been carried out yet.This paper provides a detailed account of a validation study on the Master's UEE with respect to the Messickian validity framework using the Rasch model and hence is an initial effort to provide validity evidence for Master's UEE.

According to Messick (1995), although validity is a unitary concept, six distinguishable aspects of construct validity can be highlighted as means of addressing the notion of validity as a unified concept. These six aspects of construct validity are *content*, substantive, structural, generalizability, external, and consequential.

According to Messick (1995), content aspect evidence should address the relevance and representativeness of the content upon which test items are based and the technical quality of those items. These two aspects of content validity can be threatened by *constructunderrepresentation* and *construct irrelevant* variance (Messick, 1989). The representative aspect concerns the degree to which a test is sensitive tovariations in the construct being measured (Borsboom et al., 2004). The items of a test show acceptable degree of representativeness if not only a sufficient number of items are included on the measurement instrument but also the item hierarchy map representsa sufficient spread of items with minimum gaps or overlaps. According to Baghaei (2008), items which misfitto the Rasch model are indications of constructirrelevant variance and gaps along the unidimensional continuum, representing the ability intended to be measured, are indications of construct underrepresentation.

Substantive aspect of construct validity addresses the degree to which the processes engaged by the test takers are representative of and relevant to the processes assumed in the domain under study (Messick, 1995). It adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Loevinger, 1957). The cognitive modeling of the response process of the test takers is an evidence to support the substantive aspect of construct validity.

The structural aspect of validity addresses the degree to which the scoring model matches the structure of the test (Messick, 1995). As Adam &Wu (2007, p.21) argued, "an aggregated item score is meaningful just when all the test items tap into the same latent variable. Otherwise, one outcome score for different dimensions is uninterpretable, since the same total score for students A and B could mean that student A scored high on latent variable X, and low on latent

variable Y, and vice versa for student B". Hence, the dimensionality of the test is a determining factor in the choice of its scoring model.

The generalizability aspect of validity concerns the principle of invariance which is claimed to be the essence of validity argument in the human sciences. Rasch (1960) described invariance as: "The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared" (p. 332). The group of persons who take a given test are a sample of the population of all possible test takers and the items are a sample of all possible items which could be included in the test. The item and person invariance needs due attention in generalizing the interpretation of the test scores.

The external aspect of construct validity refers to the degree to which a test is related to other tests of the same construct, tests of other constructs, or non-test behavior (Messick 1995). Evidence on external aspect can be accrued through(a)Multitrait-Multimethod (MTMM) analysis. Different measures of the same construct are expected to have a higher positive correlation than different or the same measures of different construct. (b) sensitivity to treatment and differential groups studies. Through these studies, capability of a test to detect differences between those who are supposed to have developed high levels of a construct and those who either do not possess it, or possess low levels of it, is demonstrated.

According to Messick (1995, p.746) "the consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short-and long-term". The unintended outcomes may be positive, like improving educational system, or negative, as the source of item bias or DIF.

2. Method

2.1 Participants

The participants of the current study are 18,822 B.A holders (5,184 male and 13,637 female) from mainly English Literature and Translation Studies. They sat for the English Master's UEE in 2009. This National Matriculation test screens the applicants into English Teaching, English Literature, and Translation Studies programs at M.A. level in Iran.

2.2 Instrument

The main instrument of the study is the English Master's UEE administered in 2009. Thetest is designed to measurethe applicants' general English knowledge (GE)and content knowledge. The GE part that must be taken by all candidates is an English proficiency test of 60 multiple choice (MC) items, in four areas of grammar (10items), vocabulary (20 items), cloze (10 items), and reading comprehension (20 items). The Content knowledge part consists of three separate parts with 60 MC items each. Each part is intended for the applicants to English Teaching, English Literature, or Translation Studies programs.The allotted time given to participants to complete both the GE and the content knowledge parts is 120 minutes, 60 minutes each.

3. Procedure

For the purpose of the present study WINSTEPS® Rasch software (Linacre, 2009a) version 3.68.0 and eRm 0.15-4 (Mair, Hatzinger, &Maier, 2014), an R package, were used for analyzing the data. Prior to doing the main analyses, we checked the assumptions of the Rasch model.

3.1 Rasch Model Assumptions Check

Application of item response theory (IRT) models in general and Rasch model in particular, requires observation of two assumptions: *local independence* and *unidimensionality*. Unidimensionality requires that a single dominant construct should underlie responses to the items of a test. Each item on a test may measure more than one dimension. The set of items included in a test are expected to share the same dimension the test is intended to measure. The particular dimension is expected to overwhelm the other dimensions measured by the items. Unidimensionality requires that these "other dimensions" not be shared by many items on the test and act like random noise (Linacre, 2009b). Local independence expresses unidimensionality another way. It includes but goes beyond unidimensionality. It implies that the correlation between the items of a test should be due to the single dominant dimension affecting performance on the test. After removing the effect of the dimension, the correlation should reduce to zero (Linacre, 2009b).

To check unidimensionality of the test, principal component analysis (PCA) of residuals was inspected. PCA of residuals in Table 1 shows the Rasch dimension (the intended dimension) is as big as 12.8 items (eigenvalue = 12.8) and explains 17.5% of the variation in the data.

Table 1. Dimensionality output

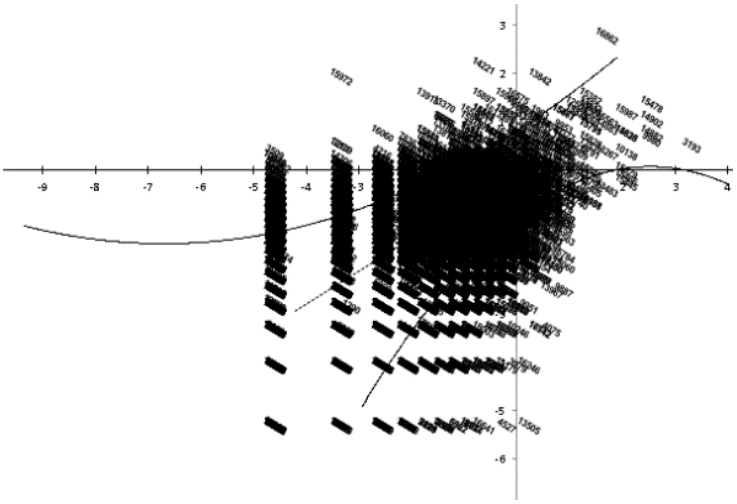
	Empirical	--	Modeled
Total raw variance in observations	72.8 100.0%		100.00%
Raw variance explained by measures	**12.8 17.5%**		**17.00%**
Raw variance explained by persons	2.0 2.8%		2.70%
Raw Variance explained by items	10.7 14.8%		14.30%
Raw unexplained variance (total)	60.0 82.5%	100.00%	83.00%
Unexplained variance in 1st contrast	**2.4 3.2%**	3.90%	

The Modeled column indicates the amount of variance explained if the data perfectly fit the data. As one can read from this column, if the data had fit the model well, 17% of the variance in the data would have been explained by the Rasch dimension. Unidimensionality of the data can be checked from the row "unexplained variance in 1st contrast" in Table 1. Two cut-offs have been suggested for the *eigenvalue* of the "other dimensions". Raich (2005) has proposed that secondary dimensions which have the strength of at least two items (eigenvalue=2) are causes of concern, whereas Smith and Miao (1994), conducting simulation studies, have

suggested that eigenvalues above 1.4 are indicators of strong-enough secondary dimensions. Judged by either criteria, the eigenvalue of 2.4 for the first contrast in the present study indicates that the test is probably multidimensional.

We followed two more steps to make sure about the dimensionality of the data. First we simulated a dataset with the same characteristics as the real dataset, which fit the Rasch model well and compared the eigenvalue of the first contrast obtained from the two data sets. The result of the analysis based on the simulated data showed an eigenvalue of 1.1 which is much smaller than the eigenvalue obtained in the present study, pointing to a relatively sizeable secondary dimension. We further explored the dimensionality of the test by conducting a more stringent analysis recommended by Wright (1977). The test was split into two halves: positively loading items on the first component in the residuals and those loading negatively. The loadings are the correlation between the items and the secondary dimension excreted from the data (Baghaei, &Cassady, 2014). Then the person measures obtained from the two halves were cross-plotted, as shown in Figure 1. In this figure, the dotted line in the middle is the Rasch dimension. Under conditions of complete fit of the data to the Rasch model, all the person measures should lie along the dotted line, but in practice it is unachievable. Standard errors of person estimates are used to build 95% confidence interval control lines (the two solid lines at the sides of the dotted line), which show limits of standard error of measurement. If the person estimates obtained from the two halves are sufficiently invariant, at least 95% of the estimates should lie between the control lines (Bond & Fox, 2007).

Figure 1. Person cross-plot of the total test



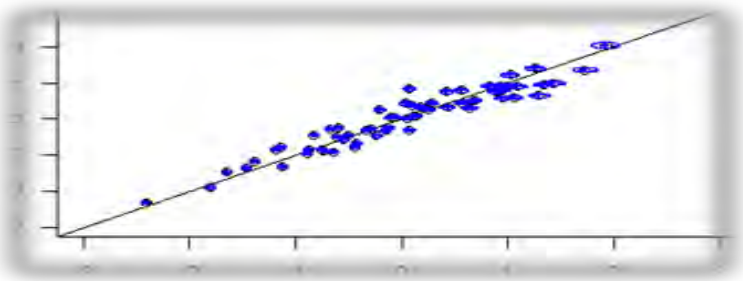
Visually inspecting Figure 1, one can observe that more than 5% of the person estimates fall outside the control bands, suggesting multidimensionality in the data.

Finally, the data were subjected to Andersen (1973) likelihood-ratio (LR) test using eRm package (Mair, et al., 2013) in R. For the purpose of the LR test, test takers are divided into two subgroups according to their mean measures and item parameters are estimated for each subgroup separately. The results of the LR testsuggest that the null hypothesis of no significance

difference between the item difficulties obtained from the two subgroups is rejected at χ^2 (59) =2557.149 ,p=000.

Figure 2 is the plot of item difficulties obtained from the first subgroup against those obtained from the second subgroup. Deviations from the diagonal line indicate lack of item difficulty invariance across the two subgroups (Mair, et al., 2013).

Figure 2. Graphical invariance check



As one can see from Figure 2, for a good number of the items even the 95 % confidence ellipses don't cross the diagonal line.

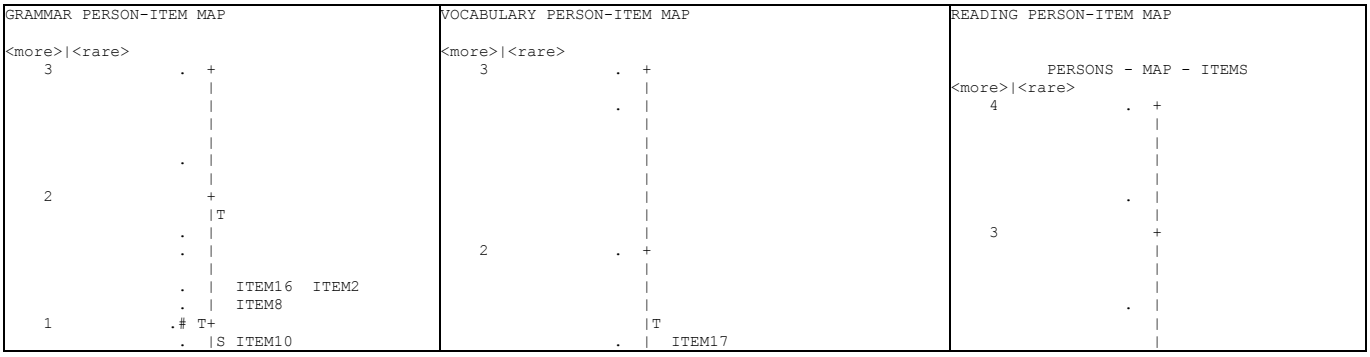
All the evidence obtained pointed to multidimensionality of the test, therefore it was decided to subject each section separately to Rasch analysis. The cloze section was not included into the analysis, because local independence assumption is most probably violated.

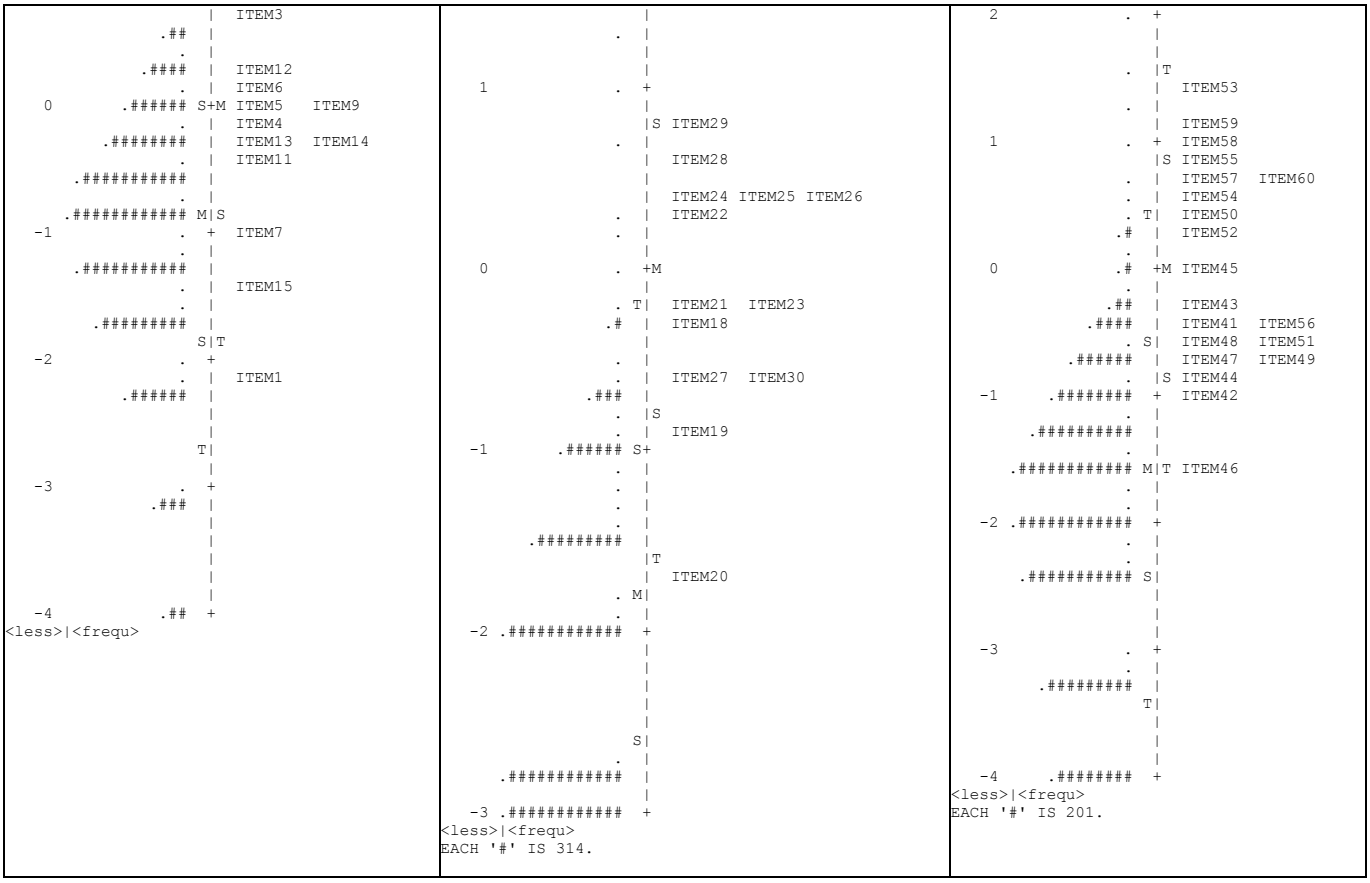
3.2 Content Aspect of Construct Validity

Evidence regarding content aspect of construct validity can be gleaned from the following sources (Wolfe & Smith, 2007): (a) inspection of person-item map: gaps and redundancies along the vertical line are causes of concern; mismatch of items' vs. persons' mean, , (b) item strata index, (c) infit and outfit statistics, and (d) point-measure correlations

The person-item map contains a wealth of arguments applicable to not only the representative aspect of Messick's content validity but nearly to the other facets as well. (Beglar,2010). Figure 3 shows the linear relationships between the Rasch calibrations for the 18,822 test-takers and 16 grammar, 14 vocabulary, and 20 reading items, respectively.

Figure 3. Person-item maps





The vertical lines represent the dimensions or the constructs which the items are supposed to define operationally. Conventionally, the mean item difficulty measures is centered at zero. On the vertical line, M represents the mean, S represents 1 standard deviation away from the mean, and T represents 2 standard deviations away from the mean. The upper part of the line locates more able persons and more difficult items,whereas the lower part represents the less able persons and less difficult items. Figure 3 shows that in each section of the test the bulk of items on the right are not well matched to the bulk of the persons on the left, indicating the test is not appropriately targeted for this group of participants. For a well-targeted test, mean ability measures of the test takers should be located around the mean difficulty of the items (around 0 logit). As one can see from Figure 3, in the grammar, vocabulary, and reading sections, mean ability of the subjects are located at 1, 2, and 2 standard deviations blow the respective mean item measures. In all the sections, the majority of the items are clustered at the top of the maps, while the spread of persons is at the bottom. This pattern is more evident in the person-item map of vocabulary and reading sections in which test takers are spread over a span of 0 to -3 and 0 to -4 logits, respectively, whereas the item difficulties span from 1.5 to -1.5 logits. Hence there are not sufficient items to estimate the knowledge of low proficiency examinees at the lower end of the scales. This evidence shows that the items in different sections of the test represent the respective content poorly.

Not only are most of the items clustered at the top of the item distributions, but also

considerable gaps exist between items of different difficulty levels in the maps. In order to have a more precise estimate of the persons who fall in these regions of ability,one needs to have more items in these areas. In the vocabulary map, there are noticeable gaps between items 17-29 and 21-22 but they are not as serious as the gaps which exist at the bottom of the line. These gaps are indicators of construct underrepresentation which threat construct validity.A more balanced test which helps differentiate between the low and high proficiencytest takers would be achieved by addition of more easy items to different sections.

Another point about the person-item maps in Figure 3 is that there are no noticeable redundancies along the line. Items with the same difficulty measures target the same level of the latent trait. According to the maps in Figure 3, there are some items that target the same level of the latent traits. These items cover the same content area and provide the same information on test taker's abilities hence are redundant. For example, items 24, 25, and 26 in the vocabulary section and items 47 and 49 in the reading section, among others, have similar difficulty estimates. Qualitative inspection of items 47 and 49 showed that both items assessed the inference making ability of the test takers under one of the passages in the reading section, hence they have the same function in measuring examinees' latent trait and contribute similar information to the test in general. Such kind of items that sample the same part of the latent trait can be removed so that the test can measure the same expanse of the latent trait but with fewer items.

Representativeness can also checked by translating the item separation index provided by the Rasch model into the item strata statistics through the following formula:$H = (4G + 1)/3$. G is the separation index which is the ratio of the true standard deviation of items / average measurement error of items. Item strata refers to the number distinct item difficulty levels or strata that the test takers' performance can define. The inspection of item strata helps ensure that a range of item difficulties have been included in the test (Smith, 2001). According to Linacre (2007a), high item strata (>2) depends chiefly on two factors: 1)item difficulty variance which is an evidence of the representativeness of the items, 2) large person sample size. The item separation statistics for grammar, vocabulary, and reading sections are 49.01, 31.02, and 37.49, respectively. When translated into item strata statistics, they become 65.68, 41.69, and 50.32, respectively. For the grammar section, for example, the strata index of 65.68 means that the test takers' performance on the grammar part of the test defined about 66 distinct levels or strata of item difficulty. Since the sample size of the current study was too large(i.e., 18,822) the measurement error was underestimated, hence the item reliability index was inflated. Since the separation index is a function of true standard deviation of itemsdivided byaverage measurement error of items, the smaller the measurement error, the more item difficulty strata with distinct distributions, which are three measurement errors apart, can be defined within the "true" distribution of the item difficulties. In the present study, the variance of the true difficulty distribution for the grammar items was .98 and measurement error for item difficulties was .02, that is the true item difficulty variance was 49 times (0.98/.02) as big as the average item difficulty measurement error, hence about 66 distributions or item difficulty strata can be defined within the distribution of item difficulties. The gaps along the hypothetical line in the person-item maps in Figure 3 belie the high separation indices. Therefore one can safely conclude that they cannot be indicative of a wide difficulty range of the items but they are artifacts of a large sample size, which are not evidence for representativeness of the items.

In conjunction with item strata and visual inspection of item-person maps, infit and outfit Mean square and point-measure correlations also provide evidence related to content validity of a test (Wolf & Smith, 2007). Items that do not fit the Rasch model may do so as a result of measuring constructs other than the one being measured by the Rasch model (multidimensionality), mis-keying, poor item quality, or local item dependence, which are all related to content aspect of construct validity. The expected value for infit and outfit mean squares is 1. Values greater than 1 indicate unpredictability or sources of variance other than the intended latent trait, whereas values smaller than 1 indicate overfit or too much predictability. The acceptable range of infit and outfit values for high stakes tests such as the present test, as suggested by Linacre (1994) is 0.80 to 1.2. Z-Standardized (ZSTD) statistics tests the statistical significance of the difference between the observed infit and outfit Mean-square and their expected value (i.e., 1). ZSTD values outside the range of -2 to +2 indicate significance difference. If the mean squares are acceptable, ZSTDs can be ignored (Linacre, 2007). ZSTDs are affected by sample size; for large sample sizes even small deviations from the expected value of mean squares (i.e., 1) tend to become statistically significant. Linacre (2007) suggests checking infit and outfit mean squares when sample size is larger than 300 because mean squares are corrected for sample size. When more than 5% of items or persons have mean squares and ZSTDs outside the acceptable range the invalidity alarm sounds (Wright&Masters, 1982; Wright & Stone, 1979). Tables 2, 3, and 4in the Appendix show the fit indices for grammar, vocabulary, and reading sections, respectively. All the mean squares are within the acceptable range, thus ZSTDs outside ±2 are ignored.

Point-measure correlation, analogous to classic item-total correlation, is the Pearson correlation between item score and the ability measure of the respondents who took the item (Smith & Wolf, 2007). Positive item-measure correlation indicates that the item scoreisconsistent with the average scores on the rest of the items, whereas negative item correlation is indicative of miskeyed items.Near zero item-measure correlations indicate that the item is a measure of another content area or construct and do not contribute to the measurement process the same thing as the other items do. No negative or near zero item-measure correlations are seen in different sections of Tables 2, 3, and 4.

3.3 The Substantive Aspect of Construct Validity

As errors or misconceptions are helpful for creating cognitively diagnostic assessments (Frederiksen, Mislevy, and Bejar, 1993; Snow and Lohman, 1989), the inspection of test takers' responses to the distractors of multiple choice items provides convincing evidence regarding substantive aspect. Three indices including the proportion of respondents who chose each distracter (i.e., p values), the average ability measure of respondents who chose each distracter (AKA "choice mean"), and the distracter-measure correlation (Wolfe & Smith, 2007b, P.209) indicate the (in)consistency of the distractor response process to the intended cognitive processing associated with the response model.

Distractors which attract less than 5% of the respondents, a rule of thumb suggested by Linacre (2007b), should be replaced by more plausible ones. About six distractors attract just about 4% of the respondents (Tables are available upon request from the first author). These distractors are inconsistent with the intended cognitive processes around which the distracters were developed, hence they are candidates for modification.

The next evidence for the substantive aspect is the mean ability estimate of all the participants who have chosen a particular option. According to Wolfe and Smith (2007) "If a distractor does not attract less able respondents then its validity as a measure of underlying construct is questionable" (p.209). Hence it is expected that the value for average measure be the highest for the keyed option. The asterisk above the correct option indicates that this expectation is not met for that specific item. This expectation is not hold in item 1 of the grammar part and item 23 of vocabulary part.So the mean ability measures of persons who got these items right are less than the mean of those who have chosen the wrong options, suggesting these distractors do not match the intended cognitive processes.

Person and item infit and outfit statistics can also provide additional evidence for the substantive aspect of construct validity. The standard mean square fit indices show how test takers' cognitive processes match the processing models developed as a part of the definition of the construct. According to Wolfe and Smith (2007), "person misfit may arise due to unmodeled examinee characteristics (e.g., guessing), specialized knowledge (e.g., test security breech or special training relating to the content of the instrument), carelessness, item bias, and response sets" (p.211). In the same vein, infit and outfit statistics outside the acceptable range of 0.80-1.20 for any given high-stakes item, indicate that the item does not collaborate with other items on the test to define the dimension being measured and the cognitive processes underlying the item are not the same as those underlying the mainstream items. As different sections of Tables 3, 4, and 5 show, all the items in all the sections fit the model, whereas more than 5% of the persons did not fit the model (person fit tables are not included in the interest of space).

3.4 The Structural Aspect of Construct Validity

The evidence, discussed in the Rasch Model Assumption Check Section, pointed to the multidimensionality of this version of the Masters' UEE. Since a single score is reported for the whole test, the structural aspect of the construct validity of the test is questionable.

We further checked dimensionality of each section separately following the steps mentioned in the Rasch Model Assumptions Checksection above: (a) Eigenvalues of the first contrasts were checked, (b) The observed eigenvalues were compared to the eigenvalue of simulated data, (c) Anderson's LR test was conducted.

As Table 5 shows, the eigenvalues of both the vocabulary and reading sections were above the 1.4 cut-off suggested by simulation studies (Smith and Miao, 1994) and below the cut-off of 2, suggested by Raiche (2005), whereas the eigenvalue for the first contrast in the grammar section was acceptable, judged by either criteria.

Table 5.Dimensionality Statistics for the Sections of the Test

Column1	Obs.Eigen	Sim.Eigen	LR P value
Grammar	1.3	1.3	0
Vocabulary	1.7	1.2	0
Reading	1.9	1.2	0

The results of the LR test showed that the item difficulties obtained from the two subgroups were significantly different at p<000. This result was corroborated by the plots of item difficulties obtained from the first subgroup against those obtained from the second subgroup. Deviations from the diagonal line indicate lack of item difficulty invariance across the two subgroups (Mair, Hatzinger, & Maier, 2013). As one can see from Figure 2, for a good number of items in all the three sections even the 95% confidence ellipses don't cross the diagonal line.

3.5 The Generalizability Aspect of Construct Validity

The Rasch modeling property of invariance is verified empirically by examining item and person calibration invariance. Item calibration invariance deals with the degree to which items measure the same dimension across different subgroups, hence have the same meaning for the subgroups, which is addressed by differential item functioning (DIF) analysis. Person calibration invariance deals with the degree to which the examinees perform differently on two tests of the same latent trait, which is the concern of differential person functioning (DPF) analysis.

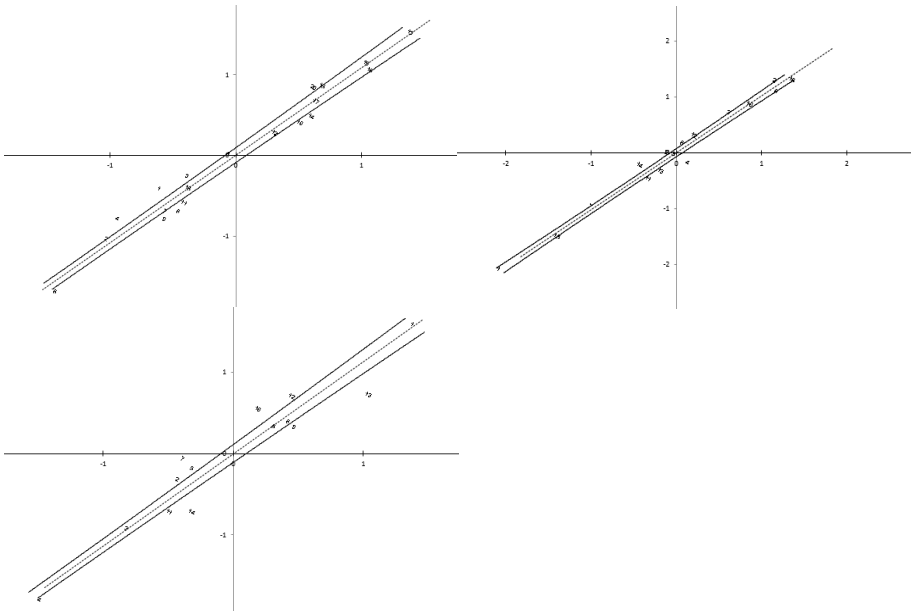
DIF in the present study was studied across gender. In DIF analysis both size and significance of the DIF contrast should be investigated. In terms of significance,the DIF values of about 62% of grammar items, 71% of vocabulary items, and 65% of reading items, are significant at P<0.05. Size refers to the degree to which item difficulty measures for the subgroups make a difference in the decisions made based on the test. Linacre (personal communication, 12, 7,2013) suggests if sum of DIF effects/number of items is two times as big as the average S.E. for person measures, then the size is significant for the group. Because of small numbers of items in each section, person ability measures havemean standard errors greater than 0.5 logits. The size of the biggest DIF of 0.39 logits (ITEM 26) is smaller than the measurement error of the person measures. Applying this criterion to the other items, we found that the DIF effect is not big enough to be considered as a threat to the fairness and generalizability of the test.

Another criterion used to judge DIF size in the present study was Draba's (1977) suggestion. Draba suggested if the difference between an item difficultyestimate for the two groups is more than 0.5 logits it should be flagged for DIF. The DIF contrast value for none of the items is more than 0.5 logits.Therefore, although the difference between the difficulty measure of a good number of items were significantly different, DIF size, judged by either Linacre's or Draba's (1977) criteria, was not substantive for the items in different sections of the test.

DIF results pointed to invariance of item calibrations across male and female samples of test takers. The same analysis was carried out by cross-plotting the item measures obtained from the two subsamples. For the purpose of the present study, test takers were split into male and female samples then item difficulties were estimated based on each sample. Figure 7 displays the cross-plots of the two sets of item calibrations for each section.

In the cross-plots in Figure 7, items which fall between the two lines have got equal item difficulty estimates (within measurement error) in the two samples.

Figure 7. Item cross-plots for the reading, grammar, and vocabulary sections

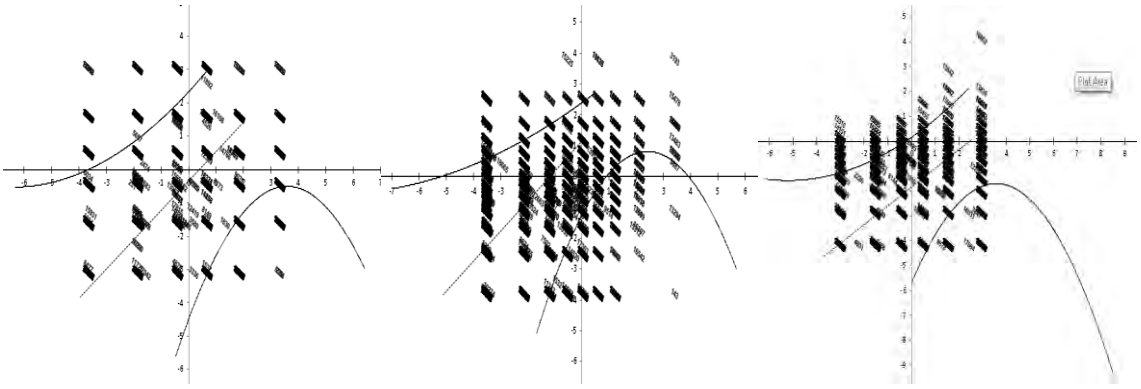


As one can see from the plots, about half of the items in each section fall beyond the control lines. On the face of it, the results of the DIF mentioned previously might seem to contradict with the plots. But there is a point worthy of note about the plots: The 95 % confidence interval lines are very close together. This is due to high statistical power of the test which reduced the measurement error and consequently resulted in control lines being close together. As is evident from the plots, most of the items that fall beyond the lines are very close to the lines. Had a smaller sample been used, the 95 % control lines would have been positioned farther from each other and more than 95% of the items would have fallen within the lines, as a result. Therefore, item measures remained invariant across the two subsamples.

Invariance of person measures also provides evidence for generalizability aspect of validity. For this purpose a test is divided into two halves based on the loading of the items on the first contrast in the residuals: items with positive loading vs. items with negative loading. Then for each person, two ability measures are estimated:difficulty measures obtained from the items with positive residual loadings and those obtained from the items with negative loadings and plotted against each other(Baghaei, 2010, 2011).

For person estimates obtained from the two halves be sufficiently invariant, at least 95% of the estimates should lie between the 95 % confidence interval lines (Bond & Fox, 2007). Figure 8 displays the cross-plots of person measures for reading, grammar, and vocabulary, sections, respectively.

Figure 8. Reading, grammar, and vocabulary person cross-plots



Visual inspection of Figure 8 shows that for the reading section more than 5% of the test takers fall beyond the control lines. But for vocabulary and grammar sections it is difficult to detect whether person measures fall enough within the control lines or not.

To clarify the point, we correlated the person measures obtained from the two sets of items: those loading positively on the first contrast vs. those loading negatively. The correlation coefficients for the two sets of person measures were .27,.32, and .27 for the reading, grammar, and vocabulary sections, respectively. These low correlations suggest lack of invariance of person measures across subsets of items.

As it is clear from the foregoing discussion item measures were invariant, whereas person measures were not. Now the question is which one is a more serious threat to validity of the test: lack of person invariance or item invariance? If item measures are not invariant, they target different ability levels for different groups, hence predictive validity of the test (in the traditional sense) is threatened but lack of invariance of person measures indicates that the two subsets of items may measure different dimensions hence construct validity of the test is threatened (J., M. Linacre, personal communication, January, 24, 2014).

In addition to invariance, representativeness of items and persons can also shed light on generalizability of a test. If the samples of items and persons are representative of all possible persons and items, the results can confidently be generalized to all the non-test situations of the same construct. Representativeness can be checked from item and person strata. For the persons and items to be representative of the universe of possible persons and items, they have to have a minimum separation value of 2. The person separation values for grammar, vocabulary, and reading sections of the test were 1.08, 0.90, and 1.07, respectively, which indicate the items of the test sections could not define at least two distinct strata of person abilities. The limited range of person abilities corroborates the result obtained by person separation index. Standard deviation of the person measures for grammar, vocabulary, and reading section of the test were 0.33, 0.53, 0.68. One reason for low standard deviation and separation might be the small number of items in each section. To test tenability of this hypothesis, the items of all the sections were put into a single analysis. The standard deviation and separation for the whole test were 0.63 and 1.5, respectively, still suggesting a limited range of person measures. Both the separation indices and the standard deviations for the separate sections indicate that the persons are not representative of the universe of possible test takers. Since the test takers are dispersed along a span of about -2 to 0, -3 to -1, and -4 to 0 for the grammar, vocabulary, and reading

sections, respectively, they represent only low ability portions of the universe of possible test takers. Unlike person separation indices, item separation indices were very high. As argued in Content Aspect of Validity section, indices of this big for items are due to large sample size,hence cannot be evidence of representativeness of the items.

3.6 The External Aspect of Construct Validity

Studies on the external aspect of construct validity examine item spread relative to the dispersion of person measures by inspecting item-person map and person strata index. To be sensitive to treatment, the item-person map for a test should have the following two characteristics (Wolfe & Smith, 2007): (a) There should be a floor effect for the distribution of the person measures and (b) the items should be widely dispersed along the line and many of the items should be located beyond the highest ability person. Messick (1989) argues that the social consequences and value implication of the test score use and interpretation is also important and the validity of it should take into consideration.

As Figure 3 shows, in the person-item map of the reading and vocabulary sections of the current Masters' UEE, most of the items are located beyond the highest person measure but the items do not cover a wide span of the respective constructs. However, in the grammar section, although the bulk of the items are slightly above the bulk of the persons, there is no floor effect for the distribution of person measures and items are not widely dispersed along the ability continuum. This indicates that if the items of these sections of the test were part of an achievement test, reading and vocabulary, to some extent, would be able to detect changes due to intervention, or differentiate between the test takers supposed to possess the constructs and those who do not possess them. In addition to visual inspection of person-item map, person strata index which represents the number of statistically different ability strata that the test can identify in a sample is another evidence towards external aspect of construct validity. Person separationindices > 2 indicate that the test is sensitive enough to distinguish between high and low proficiency test takers, hence supports the external aspect of validity of a test. Low person separation values (< 2) for different sections of the test (1.08, 0.31, and 1.07 for grammar, vocabulary, and reading, respectively) show that more items covering wider spans of the respective latent traits are needed to classify persons with different levels of the intended latent trait. Combining the evidence obtained by the spread of items and the person separation indices one can conclude that although the test has got the some potential to detect the difference between those with high and low levels of the intended constructs in the reading and vocabulary sections, lack of floor effect in the grammar section and the number of distinctly identified ability layers in all the sections limit their capability to measure test takers who vary widely in their language proficiency is limited.

3.7 The Consequential Aspect of Construct Validity

The higher the stakes of a test are, the more efforts should be put into ascertaining its consequential validity. Since performance on the UEE either makes or breaks candidates futures, its consequential validity needs due attention. The consequential validity of a test is confirmed when it provides equivalent results across differentsubgroups of test takers, and also is a pure reflection of the ability of the test takers with regard to the intended construct. For the consequential aspect of validityRasch does not provide any unique index other than those

supporting some other aspects of construct validity. Since the consequence of any test depends mostly on its degree of fit to the Rasch model and also its degree of fairness for different groups of test takers, infit and outfit statistics and DIF should be checked for the purpose of consequential validity. All the items fit the model (see Tables 2, 3, and 4) but there were considerable gaps in the person-item map of different sections of the test (see Figure3). The evidence indicated that the results may notbe based on sufficient amount of relevant information since the items were not targeted to the ability levels of most of the participants. Therefore, the result obtained and consequently the interpretations to be made based on the test, may not be reliable enough.

4. Summary of the Results and Conclusion

The results of the present study, summarized in Table6, showed that fit of the persons and items to the model, as indicated by person and item fit and point measure correlations, was acceptable, whereas the items in all the sections had targeting problems, as indicated by person item maps and item strata.

Table 6.Summary of the results

Content Aspect	
Person-item map	problematic
Item strata	Problematic (<2)
Item infit and outfit	OK
Point –measure correlation	OK
Substantive Aspect	
Distractor p value	OK
Person measure of those who chose distractors	OK
Distractor measure correlation	OK
Person infit and outfit	problematic
Item infit and outfit	OK
Structural Aspect	
Whole test	multidimensional

Grammar	unidimensional
vocabulary	unidimensional
Reading	unidimensional
Generalizability	
Invariance of item measures	OK
Invariance of item measures	problematic
Representativity	problematic
External Aspect	
Person strata	problematic
Person floor effect (grammar)	problematic
Person floor effect (vocabulary)	OK
Person floor effect (reading)	OK
Consequential Aspect	
Item infit and outfit	OK
DIF	OK
Person-item map	Problematic (noticeable gaps)

The pattern of distractor selection showed that those who had selected distractors, on average, had lower ability measures than those who had selected the correct answer. More than 5 % of the test takers misfit the model, as indicated by person infit and outfit values. The test considered as a whole was multidimensional, whereas the sections, judged by the cut-off of 2 suggested by Raiche (2005) were unidimensional. Compared with simulated data and subjected to Anderson's (1973) test the three sections were multidimensional. DIF cross-plots showed that more than 95% of the items fell between the control lines. Low Pearson correlation between the person measures obtained from the items with positive loading on the first contrast in the residuals vs. those loading negatively suggested multidimensionality of the data in all the three sections, corroborating the results obtained from the Anderson's (1973) LR test.

The person strata and person floor effect for the reading and vocabulary sections showed that the respective sections would be sensitive to treatment, had they been used to show the changes due to instruction. Finally, noticeable gaps in the person-items maps indicated that there was not enough information to estimate person abilities which might have adverse consequences

for the test takers whose futures was dependent on the results of the test. In a nutshell, the test seriously suffered from multidimensionality, low reliability, and was not targeted to the level of the test takers.

The purpose of the study was twofold: demonstrating the possibility of extending the six aspect of Messickian validity to Rasch model and exploring construct validity of the Masters' UEE test with respect to Messickian validity framework using the Rasch model. The present study showed that Rasch model can provide illuminating evidence on the six aspect of Messickian validity. The results indicated that the Masters' UEE suffers from three problems: (a) it seriously suffered from multidimensionality (b) It didn't target the test takers, and as a result (c) Its reliability was not high enough to support the unit of analysis. The difficulties of the items in all the section were much above the ability of the examinees consequently the abilities of the test takers were not estimated with enough items. Lack of appropriate targeting on the part of the items induced error into the person measures which resulted in low reliability estimates for different sections of the test. As Table 7 shows, the reliability estimates for all the three sections are well below the suggested criterion of 0.70 (DeVellis, 2003) and the mean measurement errors are relatively large.

Table 7.*Reliability Indices*

Column1	Reliability	Mean measurement error of person abilities (logit)
Reading	0.53	0.74
Grammar	0.54	1.01
Vocabulary	0.45	0.71

Although low reliability of different sections of the test is partly due to the relatively small number of items in each part, lack of appropriate targeting could be a serious threat to reliability of different sections.

Refrences

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*(1), 31-44. doi: 10.1111/j.2044-8317.1973.tb00504.x

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Baghaei, P., &Cassady, J. (2014). Validation of the Persian translation of the Cognitive Test Anxiety Scale. *Sage Open, 4*, 1-11.

Baghaei, P. (2011). *C-Test construct validation: A Rasch modeling approach*. Saarbrücken: VDM VerlagDr Müller.

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research*. Frankfurt/M.: Lang.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Baghaei, P. (2009). Understanding the Rasch model: Mashhad: Mashhad Islamic Azad University Press.

Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research, 2*(5), 1052-1060. doi: 10.4304/jltr.2.5.1052-1060

Bailey, K. M. (1999). *Washback in language testing*: Educational Testing Service Princeton,, NJ.

Barati, H., & Ahmadi, A., R. (2010). Gender-based DIF across the Subject Area: A Study of the Iranian National University Entrance Exam. *Jornal of Teaching Language Skills, 2*(3), 1-26.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118. doi: 10.1177/0265532209340194

Bond, T. G.(2003). Validity and assessment: a Rasch measurement perspective. *Metodologia de las Ciencias del Comportamiento, 5*(2), 179-194.

Bond, T., G, & Fox, C., M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Londen: Lawrence Erlbaum Associates.

Boone, W. J, & Scantlebury, K. (2005). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253-269. doi: 10.1002/sce.20106

Borsboom, D., Mellenbergh, G., J, & van H., J.. (2004). The concept of validity. *Psychological review, 111*(4), 1061-1071.

Bridgeman, B., & Wendler, C.. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology, 83*(2), 275.

Briggs, D. C, & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. United Kingdom: Sage Publications Ltd.

Committee, Scientific Advisory. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin, 3*(4), 1-4.

DeVellis, R. F. (2003). Scale development: Theory and applications. *Thousand Oaks, CA*.

Draba, RE. (1977). The identification and interpretation of item bias. *Research Memorandum*(26).

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement, 58*(3), 357-381. doi: 10.1177/0013164498058003001

Frederiksen, N., Mislevy, R., J, & Bejar, I., I. (1993). *Test theory for a new generation of tests*. New Jersey: Lawrence Erlbaum.

Kanarek, EA. (1988). Gender differences in freshman performance and their relationship to use of the SAT in admissions. Paper read at the annual meeting of the Regional Association for Institutional Research. *October, at Providence, RI*.

Linacre, J.M. (2009a). A User's Guide to Winsteps, Ministep: Rasch-Model Computer Programs. 2009. *h ttp://www. winsteps. com/winpass. htm. Last accessed, 11*, 1-286.

Linacre, J. M. (2009b). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of applied measurement, 11*, 157-169.

Linacre, J. M. (1994). *Many-facet Rasch measurement*: Mesa Press Chicago.

Loevinger, J. (1957). OBJECTIVE TESTS AS INSTRUMENTS OF PSYCHOLOGICAL THEORY: Monograph Supplement 9. *Psychological Reports, 3*(3), 635-694. doi: 10.2466/pr0.1957.3.3.635

Macdonald, P., & Paunonen, S., V. (2002). A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement, 62*(6), 921-943. doi: 10.1177/0013164402238082

Mair, P., Hatzinger, R., & Maier, M. (2014). *eRm: extended Rasch modeling*: R package version 0.15-4, URL http://CRAN. R-project. org/package= eRm.

McNamara, T. F., & Candlin, C., N. (1996). *Measuring second language performance*. Harlow, Essex, UK: Longman London.

Messick, S.. (1980a). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027. doi: 10.1037/0003-066X.35.11.1012

Messick, . (1980b). Test validity and the ethics of assessment. *American psychologist, 35*(11), 1012. doi: 10.1037/0003-066X.35.11.1012

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. *Test validity, 33*, 33-45.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5-11. doi: 10.3102/0013189x018002005

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. doi: 10.1111/j.1745-3992.1995.tb00881.x

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Razavipour, K. (2010). *National matriculation English test for English major students: It's impact and some(in)validity evidence.*, Shiraz University, Unpublished doctoral dissertation

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies, 6*(4), 141-153.

Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System, 28*(4), 541-553. doi: http://dx.doi.org/10.1016/S0346-251X(00)00037-3

Smith Jr, 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.

Smith Jr, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement*.

Smith, R. M and Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.): Objective Measurement: Theory into Practice. Volume 2. Greenwich: Ablex.

Snow, R. E. (1989). Toward Assessment of Cognitive and Conative Structures in Learning. *Educational Researcher, 18*(9), 8-14. doi: 10.3102/0013189x018009008

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement *Educational measurement (3rd ed.)* (pp. 263-331). American Council on Education: American Council on Education.

Stricker, L. J, & Rock, D. A. (2008). Factor structure of the TOEFL Internet-based test across subgroups: TOEFL iBT Research Report No. TOEFLiBT-07). Princeton, NJ: Educational Testing Service.

Tahmasbi, S., & Yamini, M.. (2012). Teachers' Interpretations and Power in a High-Stakes Test: A CLA Perspective. *English Linguistics Research, 1*(2), p53. doi: 10.5430/elr.v1n2p53

Wall, D., & Horák, T.. (2008). The impact of changes in the TOEFL® examination on teaching and learning in central and eastern Europe: Phase 2, coping with change. *Princeton, NJ: Educational Testing Service*.

Wolfe, EW, & Smith Jr, EV. (2007a). Instrument development tools and activities for measure validation using Rasch models: part I-instrument development tools. *Journal of applied measurement, 8*(1), 97-123.

Wolfe, EW, & Smith Jr, EV. (2007b). Instrument development tools and activities for measure validation using Rasch models: part II--validation activities. *Journal of Applied Measurement, 8*(2), 204.

Wright, B. D. (1977). Misunderstanding the Rasch Model. *Jurnal of Educational Measurment, 14*(2), 97-116.

Wright, B., D, Linacre, J. M., Gustafson, JE, & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement transactions, 8*(3), 370.

Wright, B., D, & Masters, Geofferey N. (1982). *Rating Scale Analysis. Rasch Measurement*: ERIC.

Wright, B., D, & Masters, Geofferey N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16*(3), 888.

Wright, B., D, & Stone, Mark H. (1979). *Best Test Design. Rasch Measurement*: ERIC.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*: Educational Measurement Solutions Melbourne.

Appendix

Table 2. Fit indices of Grammar Items

			INFIT		OUTFIT		PT-MEASURE	
Item	Measure	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.
16	1.33	0.2	1.05	2.9	1.14	4.9	.21	.26
2	1.25	0.2	1.06	4.1	1.25	8.9	.19	.26
8	1.12	0.2	.15	9.9	1.38	9.9	.13	.27
10	.87	0.2	1.00	-.1	1.02	1.1	.29	.29
3	.69	0.2	.99	-.7	1.03	1.7	.31	.31
12	.28	0.2	.97	-3.7	.96	3.0	.37	.34
6	.14	0.2	1.08	9.5	1.13	9.7	.28	.35
9	-.03	0.2	.91	-9.9	.87	-9.9	.44	.36
5	-.03	0.2	.98	-2.4	.98	-1.6	.38	.36
4	-.10	0.2	1.09	9.9	1.13	9.9	.29	.37
14	-.27	0.2	.91	-9.9	.86	-9.9	.46	.38
13	-.29	0.2	.94	-9.6	.91	-9.6	.43	.38
11	-.43	0.2	.94	-9.8	.92	-9.4	.44	.39
7	-.98	0.2	.99	-2.4	.98	-3.0	.43	.41
15	-1.46	0.2	.93	-9.9	.89	-9.9	.49	.43
1	-2.09	0.2	1.05	5.8	1.05	13.5	.40	.44

Table 3. *Fit indices of Vocabulary Items*

ITEMS	MEASURE	S.E.	INFIT	OUTFIT	PT-MEASURE

17	1.52	.04	.98	-.7	.93	-1.4	.21	.19
29	.84	.03	1.03	1.2	1.08	2.4	.23	.25
28	.62	.03	1.00	-.2	.97	-1.0	.27	.27
26	.42	.03	.95	-2.8	.87	-5.1	.33	.29
24	.41	.03	.99	-.4	.94	-2.3	.30	.29
25	.38	.03	.97	-1.6	.90	-4.1	.32	.29
22	.33	.03	.96	-2.3	.89	-4.5	.33	.30
23	-.17	.02	.93	-5.4	.86	-8.1	.41	.35
21	-.22	.02	.97	-2.5	.94	-3.7	.38	.36
18	-.34	.02	1.06	5.4	1.12	6.8	.33	.37
30	-.58	.02	1.10	9.8	1.17	9.9	.33	.40
27	-.64	.02	1.02	1.9	1.01	1.0	.39	.41
19	-.87	.02	.99	-1.6	.98	-1.8	.44	.43
20	-1.69	.02	1.04	6.2	1.05	6.5	.49	.52

Table 4. Fit indices of Reading Items

ITEMS	MEASURE	S.E.	INFIT		OUTFIT		PT-MEASURE	
53	1.49	.03	.99	-.4	.91	-2.1	.23	.22
59	1.12	.03	.96	-1.8	.85	-4.7	.29	.25
58	1.06	.03	.95	-2.5	.82	-5.7	.31	.26
55	.81	.02	.91	-5.3	.76	-9.0	.35	.28

60	.77	.02	.99	-.9	.95	-1.8	.30	.28
57	.66	.02	.95	-3.1	.86	-5.5	.34	.29
54	.52	.02	.98	-1.6	.93	-3.1	.33	.31
50	.44	.02	1.00	.1	.95	-2.3	.32	.31
52	.29	.02	.99	-1.2	.92	-3.7	.34	.33
45	.00	.02	1.07	6.6	1.15	8.3	.30	.35
43	-.29	.02	1.00	.0	1.01	.8	.38	.38
56	-.40	.02	.95	-5.5	.94	-4.6	.42	.39
41	-.47	.02	1.07	8.0	1.14	9.9	.34	.39
51	-.53	.02	.92	-9.8	.86	-9.9	.46	.40
48	-.63	.02	1.09	9.9	1.11	8.6	.35	.41
47	-.66	.02	1.03	3.9	1.02	2.0	.39	.41
49	-.73	.02	1.01	1.2	.99	-1.0	.41	.42
44	-.82	.02	.99	-1.7	.98	-2.0	.43	.42
42	-1.03	.02	1.03	5.2	1.05	5.4	.41	.44
46	-1.61	.02	1.08	9.9	1.12	9.9	.42	.48