Language Testing in Asia

RESEARCH Open Access

A Rasch-based validation of the University of Tehran English Proficiency Test (UTEPT)

Shadi Noroozi[1*] and Hossein Karami[1]

*Correspondence:
shadi.noroozi@ut.ac.ir

[1] Department of English, University of Tehran, Tehran, Iran

Abstract

Recently, psychometricians and researchers have voiced their concern over the exploration of language test items in light of Messick's validation framework. Validity has been central to test development and use; however, it has not received due attention in language tests having grave consequences for test takers. The present study sought to examine the validity of the University of Tehran English language proficiency test (UTEPT) using 6 aspects of Messick's validation framework (i.e., content, structural, external, generalizability, substantive, and consequential). To examine each validity dimension, the Rasch model was considered. To this end, the data of 394 test takers who had taken the test in 2023 were cleaned and analyzed for the assumptions of the Rasch model and then for the 6 aforementioned validity aspects. The results indicated that the structural (i.e., unidimensionality), generalizability (i.e., measurement invariance), and substantive aspects held. However, the evidence for the content, external, and consequential aspects of validity was less definitive. The implications of the findings for the validity of the test, for the realm of language testing, and for item developers and item writers will be discussed.

Keywords: Messick's validity framework, Rasch model, University of Tehran English Language Proficiency Test (UTEPT), Validity

Introduction

Test developers and psychometricians have voiced their concern over understanding and improving the psychometric qualities of language tests in recent years (e.g., Karami, 2011). Of psychometric qualities, validity has been considered not as a preference in language testing but rather as the genuine imperative (Messick, 1989), is indeed viewed as a validation process (Engelhard & Wind, 2017), and entails both empirical evidence and rational judgment (Messick, 1995). Language tests can have effects on test takers and of high-stakes and low-stakes tests; the former has more remarkable effects not only on individuals' academic careers but also on their future lives (Shohamy et al., 1996). Hence, the psychometric qualities of high-stakes tests are of paramount importance to test developers and users.

There are different approaches through which validity can be probed, and one is Messick's validation framework (Messick, 1989). Highlighting the inadequacies of criterion and content validity, Messick (1989) holds that it is construct validity that lies at the heart of validity as it is all-encompassing and embraces content, criterion, and content validity. Emphasizing the unified view towards validity, Messick (1989) proposed his progressive matrix, which has received due attention in psychometrics. Different aspects of Messick's framework can be ascertained in various ways. The Rasch model meets the conditions of true measurement and can examine whether aspects of validity hold or not. In other words, the validity-measurement nexus can be realized by considering the Rasch-Messick link (Bond & Fox, 2015). Hence, test developers and users, required to demonstrate that their tests enjoy a high degree of validity, should examine the psychometric quality of the tests they develop. The present study investigated the validity of the University of Tehran English Proficiency Test (UTEPT) through the Rasch model in light of six aspects of Messick's validity framework, explained in the paper.

Review of the literature

Validity is crucial in the development of language tests and the use of test scores (Aryadoust, 2023; Bachman, 1990). Validity involves whether the variable of interest is in essence the reason for item covariation (Devellis, 2017) and concerns demonstrating whether the interpretations and use of test scores are justified (Bachman, 1990; Messick, 1989). Denying the existence of a "valid test," Cronbach (as cited in McNamara & Roever, 2006, p.10) reminds us that it is the interpretation and use of *test scores* that are of psychometric concern. It is the process of *inductive validity* that guides us on how to move from the collected data to the respective justifications that we provide (Bond & Fox, 2015; Priest, 2000).

As high-stakes tests can have far-reaching consequences on test takers' future lives, the importance of validity is more appreciated in such tests, and validity should be ensured. Indeed, the higher the stakes of a test, the graver the consequences. Such a concern has garnered stakeholders' attention such that there is a lot of research on the validity of tests, including vocabulary size test (Beglar, 2010), multiple choice vocabulary test (Baghaee & Amrahi, 2011), listening vocabulary levels test (Ha, 2021; McLean et al., 2015), the Michigan English Test (Liu et al., 2022), TOEFL (Gu et al., 2015; Stricker & Rock, 2008), and the entrance examination held in Iran to enter universities (Alavi & Bordbar, 2020; Amirian et al., 2020; Khodi et al., 2021; Ravand & Firoozi, 2016).

Validity of Iranian Language Proficiency Tests and the UTEPT Test: Empirical studies

Several studies explored the validity of more widely known high-stakes language proficiency tests in Iran such as the English Proficiency Test (EPT) (e.g., Motallebzadeh & Khosravani, 2020); the Ministry of Science, Research and Technology (MSRT) (e.g., Khodi et al., 2024); and the Test of Language by the Iranian Measurement Organization (TOLIMO) (e.g., Heydari et al., 2014). There are also some high-stakes language proficiency tests held by top-tier universities such as the University of Tehran known as the University of Tehran English Proficiency Test (UTEPT). The language testing center (LTC) of the University of Tehran is in charge of test development and administration of the UTEPT test. This test was previously employed as a screening test for MA/MSc holders, and they did not have permission to sit for the PhD entrance examination unless they could obtain the minimum acceptable score. However, at present, obtaining the criterion proficiency score on this test is a must and prerequisite for those PhD students who want to take the comprehensive exam. Before 2023, previous versions of this test did not encompass the listening section. This test, including 100 items and with a time limit of 100 min, seems to have similar test items to the TOEFL test and is considered a crucial measure to evaluate the level of the English language proficiency of applicants. PhD students, in particular, should obtain the minimum level of score or the determined cut score as a requirement for their comprehensive exam; hence, failure in the test may change their academic journey altogether. Many universities, bodies, and institutes in our context require a score of language proficiency test that provides reliable scores and a measure that has the potential to provide valid interpretations of scores. The UTEPT test is claimed to meet the psychometrics quality, and the provided scores are valid for 2 years.

The UTEPT test itself and the data obtained from such a test have been examined for various purposes such as comparison of various differential item functioning (DIF) techniques (e.g., Fidalgo et al., 2014; Karami & Khodi, 2021). Validation of such a test itself is pressing, and the validity of such a crucial test has been investigated in the last decade (e.g., Alavi et al., 2011; Amirian et al., 2014; Karami, 2011, 2013 Rezaee & Salehi, 2009; Rezaee & Shabani, 2010).

Fairness and justice are at the heart of validity (Randall et al., 2024), and measurement invariance can be investigated through various DIF techniques to ensure validity (McNamara & Roever, 2006). Several studies scrutinized the DIF of various versions of the UTEPT test. As an example, Karami (2013) explored gender differential performance on the UTEPT test using generalizability theory, and the results reflected that the test was dependable and free of gender bias. In another research, Karami (2011) examined the measurement invariance of a version of the UTEPT test across genders using the Rasch model, and it was shown that test items did not favor any gender over the other. Furthermore, Rezaee and Shabai (2010) used logistic regression (LR), and Amirian et al. (2014) employed LR and Mantel-Haenszel (MH) methods to explore the existence of bias of test items towards either gender. In the former study, it was concluded that 39 items out of 100 items showed “negligible” DIF. In the latter, the results indicated that 28% of the data showed “negligible” DIF. Also, a study by Alavi et al. (2012) investigated measurement invariance across academic fields via LR and MH methods, and the test was shown to be free of bias. In another study by Salehi and Tayebi (2012), measurement invariance across gender of the reading section was probed through 3 steps of logistic regression, and it was concluded that the items of the reading section did not favor either gender.

A series of studies examined the construct validity of different versions of the UTEPT test through different approaches. In Rezaee and Salehi’s (2009) study, the construct validity of the grammar and vocabulary sections of UTEPT were explored through the multitrait-multimethod (MTMM) approach, and it was found that the test enjoyed discriminant and convergent validity. Salehi (2011), in another research, conducted exploratory factor analysis (EFA) to uncover the factor structure of the reading section, and the results displayed that there was under-representation in this section. In another study by Salehi (2012), the construct validity of a version of the UTEPT test was investigated through a triangulation of approaches of MTMM, EFA, and intersubject correlations. His findings revealed that the test enjoyed construct validity.

All in all, there has been a wave of research on the validity of this test during the past two decades (and they differ in considering the section of the test (e.g., reading, grammar, vocabulary), the data analyses performed (e.g., exploratory factor, multitrait-multimethod, different DIF techniques), and the aspects of validity (e.g., generalizability across gender or academic field of study) that they scrutinized). Furthermore, due to such limitations in using MH as being sensitive to sample sizes (Alavi et al., 2012) or such limitations in using MTMM as inapplicability to the reading section (Rezaee & Salehi, 2009), the construct validity cannot be comprehensively and accurately probed using these approaches. In addition, several studies (e.g., Alavi et al., 2012; Rezaee & Shabai, 2010; Salehi, 2011) have suggested the usage of item response theory (IRT) to explore measurement invariance and construct validity as they can provide more information. In Karami's (2011) study, the Rasch model was used but only DIF across gender was examined, and other aspects of construct validity in light of Messick's framework were not explored.

While some studies have ascertained this test's construct validity, the examination of the validity of the new versions of the UTEPT test including the listening section through the Rasch model considering Messick's validation framework (1989) has hitherto not received due consideration by psychometricians. Furthermore, there has been no empirical evidence that can espouse the construct validity of the whole test considering this newly added section.

Rasch model and Messick's framework validity

As for the importance of the Rasch model, the validity-measurement nexus can be realized by considering the Rasch-Messick link (Bond & Fox, 2015). It has been argued that the Rasch model approach towards validity is so comprehensive, a view that conjoins considerations for content, criteria, and consequences under the umbrella term of construct framework to test hypotheses concerning the meaning and use of test scores (Wolfe & Smith, 2007). Further, having a tight relationship with fundamental measurement and meeting the conditions of conjoint measurement, the Rasch model is a true definition of measurement unlike item response theory two- or three-parameter logistic models, and hence, ignorance of such a true definition of measurement may render our interpretations of test scores inadequate. The Rasch model also plays a supervisory role in considering how the validation process operates at the interface between the development of a measure and the collected data (Bond & Fox, 2015; Boone et al., 2014; Michell, 2004).

In a research study, Baghaee and Amrahi (2011) probed the six aspects of the Messikian validity framework of a multiple-choice vocabulary test with the Rasch model, and the findings revealed that the Rasch model can provide insights into the validity of language tests. Ravand and Firoozi (2016), in another research, investigated the construct validity framework of language test items of the Iranian university entrance examination for MA in English majors. The results indicated that some indexes supported aspects of the validity framework and some did not; evidence for supporting aspects of the validity of the test was not definitive. While some research studies (e.g., Baghaee & Amrahi, 2011; Ravand & Firoozi, 2016) have been carried out on test validation using the Rasch model in light of comprehensive Messick's matrix, no single study exists to examine the validity of the UTEPT test that includes the newly added section of listening section in our context. Explanations for the Rasch model and Messick's (1989) framework are in order.

As per suggestions by Messick (1989, 1995), construct validity subsumes six distinct aspects that function as standard validity criteria applied to the measurement, either in education or psychology. These six aspects involve content, substantive, structural, generalizability, external, and consequential. All aspects are examined in the current study. Each of the aspects will be explicated one by one.

Regarding content validity, Messick (1995) explains that this aspect deals with evidence collected on relevance, representativeness, and technical quality. McNamara and Roever (2006) highlighted Messick's (1989) *construct underrepresentation* and *construct-irrelevant variance*. As for the former, a measure may require less of the examinee than is intended and the assessment would be too narrow, and regarding the latter, the variances observed in scores might be because of factors other than the ability of test takers or in other words, the assessment is too broad. Construct underrepresentation is when an assessment cannot embrace intended facets of the construct and can be examined with item separation strata (Wright & Masters, 2002) and the Wright map provided by the Rasch model via the Winsteps software (Linacre, 2013). Construct irrelevant variance, on the other hand, concerns when the assessment is tapping other constructs or other reliable variances corresponding to other facets and can be checked via fit statistics.

The substantive aspect involves the extent to which "theoretical rationales relating to both item content and cognitive processing models adequately explain the observed consistencies among item responses" (Wolfe & Smith, 2007, p. 207). As the definition is speaking, it can also be viewed as a complementary for content validity. Person fit statistics (i.e., infit and outfit mean squares) are used to examine this aspect (Wolfe & Smith, 2007).

The structural aspect involves whether the theory of construct domain can account for the scoring structure or model (Messick, 1995). A single or unidimensional construct underlying a set of items is one of the Rasch model requirements. Unidimensionality, as a hallmark of true scientific measurement, indicates the Rasch model's focus on fundamental measurement. Also, the Rasch model underscores measuring one attribute of an object at a time, even in the case of complex measurement situations. Unidemensionality is examined by running principal component analysis on residuals that can reveal the presence of probable factors or dimensions through decomposing correlation matrixes of items and persons (Bond & Fox, 2015; Linacre, 1998).

According to Cook and Campbell (as cited in Messick, 1995, p.6), generalizability "examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks." This aspect of validity is checked via differential item functioning (DIF) techniques, and in this study, DIF is computed through a *t*-test approach (Wright & Stone, 1979) to check for item calibration invariance. Furthermore, invariance is checked by splitting the items into two subtests based on their positive and negative residual loadings, estimating person ability estimates for items with negative and positive residual loadings, and finally running correlations between the ability estimates. Large correlations can be indicative of invariance.

As for external validity, this aspect of validity scrutinizes the dispersion of difficulty estimates relative to the spread of ability estimates (Wolfe & Smith, 2007). The person strata index can provide pieces of evidence for this aspect of validity. Last but not least, the consequential aspect of validity is concerned with the value implications of test score interpretation as a source on which decisions, actual and potential consequences depend. In this regard, evidence should be collected for any sources of invalidity from bias, through fairness, to distributive justice. As per suggestions by Messick (1995), no adverse consequence should derive from any sources of construct underrepresentation or construct irrelevant variance that can negatively impact test takers' performance.

Method

Dataset

The participants of the current study ($N = 394$), including 133 males and 261 females, consist of examinees who sat for one of the University of Tehran English Proficiency Test (UTEPT) held in 2023; the total number of examinees that year was approximately 6000. No other demographic information such as their age or academic field of study was available. They mainly included PhD candidates who had wanted to show their acceptable level of general English proficiency before their comprehensive exam and MA graduates willing to render their scores to the committee members at their PhD interviews.

University of Tehran English Proficiency Test (UTEPT)

The instrument for the current study was the University of Tehran English Proficiency Test (UTEPT) administered every other month by the English department. Candidates' right level of English proficiency can be revealed by test takers' performance on the general English (GE) section of the test. This test has 100 items altogether and encompasses four sections: structure and written expressions (30 items), vocabulary (30 items), reading comprehension (25 items), and listening comprehension (15 items). The format of the questions is multiple-choice. The reading comprehension section consists of 5 passages with 5 questions. The listening section comprises 2 lectures, and the number of items in each lecture differs (1 of them includes 7 items and the other 8 items). The amount of time allocated to completing the test is 100 min.

Data analysis

The initial step for analyzing the data was to clean them by discarding the data of the examinees who had obtained a total score of zero or endorsed all the items correctly. The rationale behind this action was that the parameters are not estimated for those who score 0 or get the maximum score. Also, each section of the test was examined for cleaning and was analyzed separately on SPSS. It should be noted that the reliability of the test was also inspected (Cronbach $\alpha = .89$), and it seems that the test is of high reliability, which is a pre-requirement for validity (Bachman, 1990). Prior to doing further analyses, the whole data were imported to the Winsteps software (Linacre, 2013). The reason was to check the model-data fit and the assumptions of the Rasch model (i.e., unidimensionality and local independence). Unidimensionality is concerned with focusing on and measuring a single underlying attribute or dimension at a time (Bond & Fox, 2015). Examining the unidimensionality assumption is empirically continued by running principal component analysis (PCA) on the data to check for a single dominant factor that can explain response patterns. To this end, Winsteps provides a table of standardized residual variance that is composed of two components: *raw variance explained by measures* and *raw unexplained variance*. The former corresponds to the amount of variance that the Rasch dimension can explain. On the other hand, the latter is concerned with the variance not explained by the Rasch dimension. This unexplained variance is accounted for by other activity that is pertinent to residuals; i.e., random noise and off-dimensional item-correlated activity. In essence, the residuals should demonstrate no structure, and hence, the unexplained variance should show no departure from the Rasch specifications/criteria. Linacre (2021) argues that a perfect unidimensionality can never be met. Hence, the question of unidimensionality must be put as "Is the lack of unidimensionality in my data sufficiently large to threaten the validity of my results?" (Linacre, 2021, p.589). The aforementioned assumptions will be examined in the results section.

Results

Rasch model and its assumptions

To check for the Rasch model assumptions, in general, and dimensionality, in particular, a principal component analysis (PCA) was run on the whole test and the results indicated that the Rasch dimension accounted for 21.9 or 30 eigenvalues or items, amounting to 17.9% of the observed variance in the data. However, the first 3 contrasts explained 3.5, 3, and 2.7 eigenvalues (2.9 %, 2.5%, and 2.2% of the observed variance, respectively). The strength of the first contrast that includes the lion's share of the unexplained variance was above 3; the presence of another dimension is probable although 3 items out of 100 items do not seem to be suggestive of a secondary dimension. We cross-plotted the results of the total test with all other subtests and among the subtests themselves and it was observed that the unidimensionality of each section should be examined separately. Because more than 5% of the ability estimates fell out of the identity line, the unidimensionality of the whole test seems to be dubious (Linacre, 2021). As such, the unidimensionality of each section was examined separately, and it was shown that this assumption holds; a detailed explanation is provided in the structural aspect of validity. Local independence for each section of the test was checked, and all sections except for the listening section revealed that this assumption holds.

Content aspect of construct validity

The content aspect deals with the "specification of the boundaries and structure of the construct domain" (Messick, 1995, p. 745). The content aspect of validity concerns collecting evidence for content relevance, representativeness, and technical quality (Messick 1989, 1995).

Representativeness

Representativeness is concerned with whether the test covers the domain content (Messick, 1989). Regarding UTEPT, no information on the test specification or the constructs used for the writing of test items is available. Hence, the item-person map in Winsteps, known as the Wright map, that spreads both ability measures and difficulty estimates on the same scale can provide insights (Boone et al., 2014). Figure 1

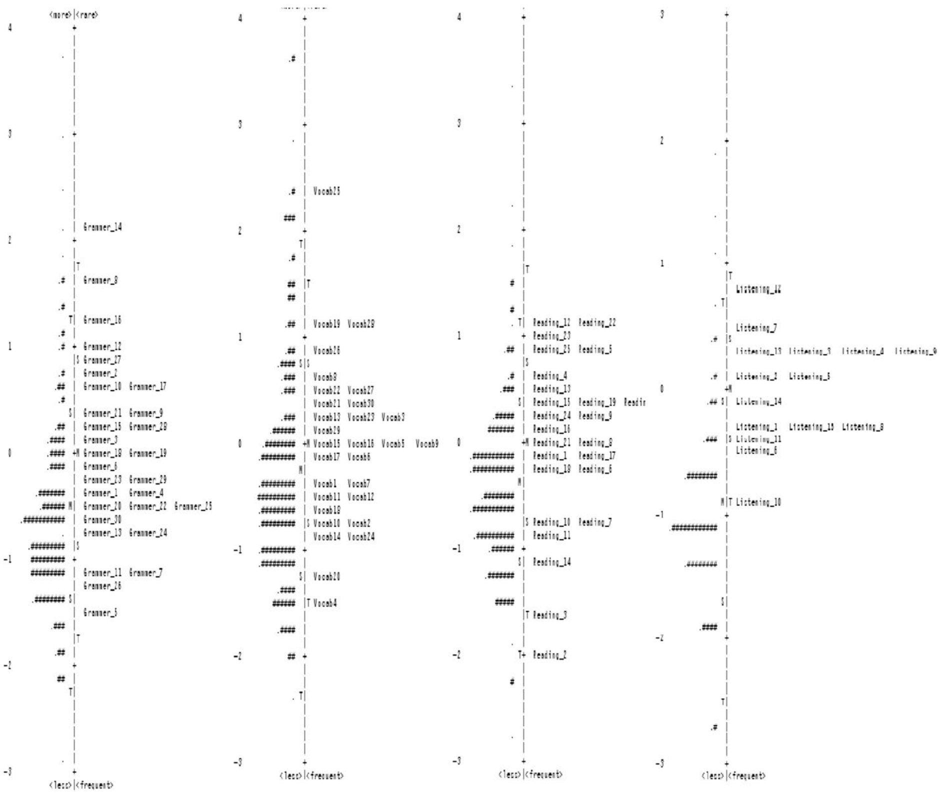

Fig. 1 Wight maps of all sections

indicates the relative location of ability measures and difficulty estimates of 30 grammar, 30 vocabulary, 25 reading, and 15 listening test items.

As evident in Fig. 1, the upper part of the vertical line shows the location of persons with higher ability levels and items with larger difficulty, whereas the lower part is indicative of less able persons and easier items to endorse. As can be seen, the average of item difficulties is centered at 0. Boone et al. (2014) emphasized that the precision of measurement is dependent upon to what extent the means of difficulty estimates and person abilities are close to each other. Figure 1 indicates that the mean of ability estimates for grammar, vocabulary, reading comprehension, and listening sections are located at - 0.53, -0.25, -0.39, and -0.94, respectively. Only the mean of person ability estimates in the listening section was located at two standard deviations below its respective mean of item difficulty estimates. In all sections except for listening, it seems that the majority of test items and persons' measures are clustered in the center of the map. In the listening section, the majority of person measures are clustered at the bottom ranging from about 0.5 logits to -2 logits, showing that listening items seem to be difficult, and do not cover a wide range of ability estimates, especially the low-ability persons. The listening section includes the highest/largest number of redundant items: Items 13, 3, 4, and 9 seem to be measuring the same construct. As for the pronounced gaps, the listening and reading sections appear to have the largest gaps or major gaps between items 10 and 6 and items 7 and 18, respectively.

Item strata and separation can be used as indicators for distinguishing groups of items and can examine representativeness. Item separation index, G_I, reflects the spread of items on the measured variable. Item strata are calculated through this formula: Strata $= (4G_{\text{item}} + 1)/3$, in which item separation index (G_{item}) is estimated as the ratio of the adjusted item standard deviation over the average measurement error, the variance not explained by the Rasch model. Person and item strata indicators should be over two to show the measure can separate items and persons into two distinct groups (Wolfe & Smith, 2007). The separation index for the grammar, vocabulary, reading, and listening sections were 6.83, 6.31, 6.71, and 3.41, respectively, meaning that the reading section, for instance, was able to differentiate between about seven groups of ability levels. Their respective item strata were as follows: 9.44, 8.73, 9.28, and 4.88, indicating that grammar item difficulties, for example, can define more than nine statistically distinct levels. The results obtained from this index are in disagreement with those obtained from the Wright map.

Technical quality

Technical quality involves indices that are concerned about and are made use of to examine the quality of any item (Wolfe & Smith, 2007). Infit and outfit mean square (MNSQ) statistics are used to investigate technical quality (Bond & Fox, 2015). Their respective *Z*-statistics (i.e., ZSTD infit and ZSTD outfit), or standardized values, show whether the infit or outfit MNSQ statistics are significant or not (Linacre, 2021). In the present study, infit and outfit MNSQ values falling in the range of 0.7 to 1.3 reflect an acceptable range of fit (Bond & Fox, 2015). Values above + 1.3 show noise in the data and hence underfit, which means some persons used special knowledge, guessing, and so on. Also, values less than 0.7, suggest overfit, or Guttman-like response strings; in other words, this index shows little variation and the response pattern seems to be too deterministic. The Rasch model is indeed probabilistic or *stochastic* rather than deterministic (Bond & Fox, 2015).

Outfit MNSQ statistics all fell in the acceptable range except for grammar items 14, 8, and 3 (i.e., 1.85, 1.34, and 1.31, respectively) and vocabulary items 28 and 21 (i.e., 1.55 and 1.51, respectively). All infit MNSQ statistics fell in the acceptable range except for the aforementioned vocabulary items 28 and 21 (i.e., 1.31 and 1.34, respectively). The item characteristic curves of these items were checked to see whether items had a radical departure from their expected curves, the ones predicted by the Rasch model (Fig. 2). Grammar item 14 and vocabulary item 28 had a very radical departure from their expected curves, as a result, they were removed from further analysis of the data. We cross-plotted the ability estimates that had been calculated once with the inclusion of misfitting items and another time without their inclusion, and it was observed that ability estimates were comparable. As per suggestions by Wright and Masters (1982), to be on the safe side, no less than 5% of the items should misfit. It seems that validity is not under question.

Furthermore, technical quality can also be examined by point-measure correlations, which is the Pearson correlation between scores on specific items and the rest of the test items. According to Linacre (2021), negative or near 0 value reflects that the item is misfitting. Zero values show that the item is either too difficult or too easy. Through

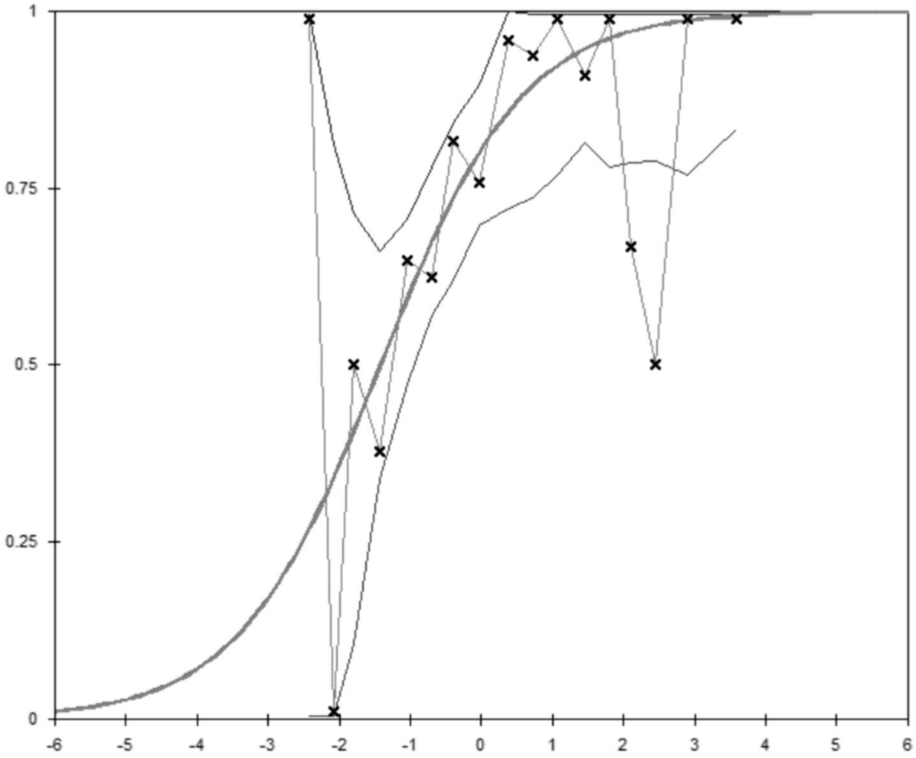

Fig. 2 Item characteristic curve for grammar item 14

the inspection of item-point measure correlation (i.e., the observed correlation), grammar items 14, 8, and 3 had correlations of 0.00, 0.24, and 0.01, respectively, and their expected correlations were 0.30, 0.33, and 0.37 respectively. For example, grammar item 14 appeared to be too difficult and the majority of the test takers could not endorse the item. Vocabulary items 28 and 21 enjoyed the small, observed correlations of 0.14 and 0.11 compared to their large expected correlations (i.e., 0.42). All the results indicate that there was not a close match between these items' observed and their expected correlations, and hence, their behavior was not in line with the rest of the items. The findings are in agreement with their fit MNSQ statistics, corroborating their misfit.

Structural aspect of construct validity

When a test measures only one construct at a time, it can be said that the instrument or test is unidimensional. To examine unidimensionality, Winsteps runs a principal component analysis (PCA) on the residuals. As shown in Table 1, a PCA on the grammar section showed that the Rasch dimension explained 8.5 eigenvalues or 22.1% of the observed variance in the data. The first contrast explained the small amount of 1.8 eigenvalues or 4.5% of the total amount of observed variance, indicating that the detected residuals not accounted for by the Rasch dimension are random or due to chance alone. Stated more precisely, the Rasch dimension appears to explain five times more variance compared with the variance that the first contrast accounts for. Further, a PCA on vocabulary, reading comprehension, and listening comprehension sections unraveled that the Rasch dimension accounted for 9.3, 6.4,

Table 1 Standardized residual variance (in eigenvalue units)

	Total raw variance in observations (percentage)	Raw variance explained by measures (percentage)	Raw variance explained by persons (percentage)	Raw variance explained by items (percentage)	Raw unexplained variance (total)	Unexplained variance in 1st contrast (percentage)
Grammar	38.5 (100.0%)	8.5 (22.1%)	2.5 (6.5%)	6.0 (15.6%)	30.0 (77.9%)	1.8 (4.5%)
Vocabulary	39.3 (100.0%)	9.3 (23.7%)	3.4 (8.8%)	5.9 (14.9%)	30.0 (76.3%)	1.9 (4.8%)
Reading	31.4 (100.0%)	6.4 (20.4%)	1.8 (5.8%)	4.6 (14.7%)	25.0 (79.6%)	1.8 (5.7%)
Listening	17.5 (100.0%)	2.5 (14.2%)	0.8 (4.3%)	1.7 (9.9%)	15.0 (85.8%)	2.8 (15.8%)

and 2.5 eigenvalues, respectively. Regarding the vocabulary section, the first component accounted for 1.9 eigenvalues or 4.8% of the total observed variance, reflecting that the residuals are random as the value. As for the reading section, the first component explained 1.8 eigenvalues or 5.8% of the total observed variance, indicating that the pattern of residuals is random. With respect to the listening section, the first component accounted for 2.8 eigenvalues or 15.8% of the total observed variance. To further examine the presence of another dimension in the listening section, we cross-plotted person ability estimates obtained using the negatively loaded items on the first contrast and positively loaded items on the first contrast. It was observed that the majority of the items fell in the 95% confidence interval, confirming the unidimensionality of this section. Hence, unidimensionality seems to be safely supported.

External aspect of construct validity

This aspect of validity examines the spread of item difficulties relative to ability measures (Wolfe & Smith, 2007). To explore this aspect, person strata and person separation index can be used. Person strata can show measurably distinct groups of persons using the following formula: person strata $= (4G_p + 1)/3$, where G_p is Rasch person separation. The person separation index is employed to estimate or predict the distribution of persons on the construct and is calculated through the following formula: $G_p = SA_p/SE_p$, where SA_p represents the adjusted person standard deviation and SE_p, the average measurement error. As per suggestions by Bond and Fox (2015), person strata should be used when the distribution is the result of very high and very low person abilities whereas person separation can be employed when the sample is large and normally distributed. As the ability estimates ranged from very low to very high in our distributions, we reported the person strata of grammar, vocabulary, reading, and listening sections as follows: 2.57, 3.09, 2.05, and 1.30, respectively. It can be concluded that the vocabulary section was able to distinguish between at least three levels of abilities while the grammar and reading sections were able to distinguish between at least two levels. However, the listening section failed to distinguish between at least two levels and the probable reason may be its small number of items or the items were so difficult that the examinees all found them challenging. Hence, the external validity of the test cannot be fully and safely supported and test designers should consider this issue for future versions of the test.

Generalizability aspect of validity

As Messick (1989, p.56) nicely put it: "The extent to which a measure's construct interpretation empirically generalizes to other population groups is here called *population generalizability* and to other tasks representative of operations called for in the particular domain of reference, *task generalizability*" (italics in the original). This aspect involves the degree to which an instrument can maintain its meaning and interpretability across subgroups, say, gender (Englehard & Wind, 2017). Measurement invariance of item measures across genders was examined by running differential item functioning (DIF) analysis to check whether test takers belonging to either group and with the same level of ability have the same probability of getting an item right. The amount of DIF is computed by a separate calibration t-test approach (Wright & Stone, 1979).

The most important column provided by Winsteps is DIF contrast and the test of its significance, Rasch-Welch probability. According to Zwick et al. (1999), there are three DIF categories: A, B, and C, having negligible, slight to moderate, and moderate to large DIF, respectively. Items having slight to moderate and moderate to large DIF are displayed in Table 2. The reason why merely these are included in the table is that items showing negligible DIF do not jeopardize the validity of the test and do not render the test biased (McNamara & Roever, 2006). It should be noted that items flagged as showing slight to moderate and moderate to high are displayed in Table 2. The column of DIF contrast is of interest as it is the difference between DIF sizes; $|\text{DIF contrast}| \geq 0.42$ logits indicate slight to moderate DIF, and $|\text{DIF contrast}| \geq 0.64$ logits reflect moderate to large DIF. The next column is t-statistic or a two-sided test for the difference between the means based on the standard error of means (Linacre, 2021), which equals DIF Contrast over Joint S.E. Furthermore, Rasch-Welch probability is the test of significance of the probability of the t value or the differences to reject the null hypothesis at $p < 0.05$.

In the current study, females were coded as 1 and males as 0. Hence, items with positive DIF contrast were in favor of females and items with negative DIF contrast were against them. As displayed in Table 2, out of 12 items, three items had moderate to large DIF: vocabulary item 22 seemed to favor females; however, listening items 4 and 12, appeared to favor males. The other nine items had only slight to moderate DIF, and DIF

Table 2 DIF results

Number	Item	DIF contrast	*t*	Rasch-Welchprobability
1	Grammar 4*	0.53	2.30	0.0219
2	Grammar 5*	− 0.53	− 2.19	0.0290
3	Grammar 11	0.45	1.87	0.0624
4	Grammar 26	0.44	1.80	0.0723
5	Vocabulary 7*	− 0.62	− 2.64	0.0088
6	Vocabulary 22*	0.74	3.06	0.0024
7	Vocabulary 24	0.44	1.83	0.0676
8	Vocabulary 30*	− 0.60	− 2.37	0.0183
9	Reading 5*	− 0.56	− 2.02	0.0442
10	Reading 10*	0.50	2.13	0.0341
11	Listening 4*	− 0.66	− 2.33	0.0206
12	Listening 12*	− 0.74	− 2.27	0.0240

* $p < 0.05$

contrasts of only six items turned out to be significant at $p < 0.05$ as shown in Table 2 by asterisks. As McNamara & Roever (2006) nicely put it: "Differential item functioning is a necessary but not sufficient condition for bias" (p. 83), further investigation is required. Then, items that were flagged as having DIF were analyzed to uncover the reasoning behind their DIF; their contents were checked and none of the items included gender-related clues. There were only three items with moderate to large DIF and their contents seemed not to favor any genders.

Further, to strengthen our reasoning and ensure measurement invariance, a stricter test suggested by Linacre (2021) was done in the current study by splitting items of each section into two subtests. This division was done based on the positive and negative item residual loadings provided by the Winsteps software. The process was such that person measures were first estimated using the negatively loaded items on the first contrast and next using positively loaded items on the first contrast. The Pearson correlations between the ability measures for grammar, vocabulary, reading comprehension, and listening sections were $r = 0.47$ (disattenuated correlation = 1), $r = 0.53$, $r = 0.27$, and $r = -0.21$, respectively. The largest correlation belonged to the vocabulary section and the lowest to the listening section having a very small negative correlation. It should be noted that the reason why the correlations were low may be due to the small number of items, leading to the restriction of range which influences the size of correlations, especially in the listening section. Notwithstanding their low coefficients, it may seem that correlation coefficients show invariance, confirming that items were not flagged as having DIF.

Substantive aspect of validity

The substantive aspect of validity deals with the degree to which theoretical rationales related to item content and cognitive processing can account for the consistencies among response items (Wolfe & Smith, 2007). Person fit statistics can reflect whether the patterns of response strings predicted by the Rasch model are in agreement with the empirical evidence or observed response strings. Linacre (2021) has advised to consider infit and outfit statistics before checking their respective standardized statistics (ZSTD). Hence, fit MNSQ statistics were checked to see if they were above 1.5 or below 0.5, which may have different causes from lucky guessing, through carelessness, to special knowledge (Bond et al., 2020). Person-infit MNSQ statistics of each section were also examined to check whether they were above 1.5. No misfitting person was observed.

Consequential aspect of validity

As stated earlier, this aspect of validity deals with value implications of score interpretation as a basis for future decisions and consequences, and hence, it is critical. A particular index is not provided by the Rasch model as for this aspect and other indexes provided to support other aspects of validity can be made use of. The inspection of person-item map and person and item fit statistics can be helpful as the extent to which items and persons fit the model can provide evidence to ensure consequential validity. Out of 100 items, 5 items (i.e., grammar items 14, 8, and 3 and vocabulary items 28 and 21) showed misfit, and there were no misfitting persons. However, there are some pronounced gaps and redundancies in the map, which shows the measure may not be considered a dependable instrument on which stakeholders can base their decisions and all person ability levels are not targeted. Hence, the consequential validity of the test appears to be under question although there were no person misfits.

Discussion and conclusion

To the best of our knowledge, this study is the first attempt to investigate the validity of the University of Tehran English Proficiency Test (UTEPT) in light of Messick's framework using the Rasch model. six aspects of content, structural, external, and generalizability validity were examined.

It was concluded that the items of this test seem not to fully meet the criteria for a good measurement, that is all aspects of validity were safely supported except for the content aspect and external aspect. As for the content validity, the test items seem not to be representative and cover the construct comprehensively. Also, there are pronounced, large gaps. This finding of under-representation seems to be in line with those of some studies (e.g., Ravand & Firoozi, 2016; Salehi, 2011), solidifying the need for more comprehensive and representative test items. It should be noted that in Ravand and Firoozi's (2016) study the measure included language test items of the Iranian University Entrance Examination (IUEE) for MA majors, and thus, the tests may not exactly be comparable. On the contrary, our findings do not mirror those of Beglar (2010) and those of Baghaee and Amrahi's (2011) research study, and the reason may be the nature of the tests themselves because vocabulary items were rather large in both studies and hence a wide range of ability estimates was targeted. It seems that in high-stakes tests such as the UTEPT test and IUEE, the number of items in each section is not large, and representativeness and comprehensiveness cannot be met.

Speaking of the structural aspect of validity, each section, in the current study, tapped a unidimensional construct predicted by the Rasch model. This result is consistent with that of other studies (e.g., Beglar, 2010; Ravand & Firoozi, 2016), suggesting that each section of the test taps one underlying trait at a time. Although our findings revealed that the whole test seemed to be multidimensional similar to the result of Ravand and Firoozi's (2016) study, this multidimensionality does not jeopardize the structural aspect of the validity, which shows the whole test consisted of different sections and taps various dimensions but each section is tapping merely one construct at a time.

With regard to the external aspect of validity, all sections except for the listening section distinguished between at least two levels of language proficiency. As for the listening section, a plausible explanation might be they were challenging items or their numbers were small. This result partially corroborates that of some studies (e.g., Ravand & Firoozi, 2016) as they concluded that while the indexes showed that external validity could not be fully supported, the test had the potential to differentiate between different proficiency levels. Our finding seems to be contrary to that of some research studies (e.g., Baghaee & Amrahi, 2011; Beglar, 2010) as their measures encompassed a large number of items and probably had the potential to target and differentiate between a wide range of ability levels.

Investigating the generalizability aspect revealed that no item showed differential item functioning across genders. This result aligns with that of other DIF studies (e.g., Karami, 2011, 2013; Salehi & Tayebi, 2012) such that the UTEPT test does not favor either gender, indicating test developers and item writers are meticulous not to include gender-related words in different sections and themes or topics in the reading comprehension section. Furthermore, the listening section, as a newly added section, seems to remain bias-free when it comes to gender-related issues. The finding, in contrast, is in partial disagreement with that of Ravand and Firoozi (2016), the reason being that their indexes employed for examining this aspect of validity might have been sensitive to the number of test items or sample size, for instance.

Examining the substantive aspect of validity unraveled that the response patterns observed were in line with those predicted by the model. This finding seems to support that of Ravand & Firoozi (2016), suggesting that the response processes of testees fail to align with those processes considered by item developers and that the individuals who had taken the test might have resorted to guessing, for instance. Although the entrance examination penalizes for incorrect answers, resorting to guessing and having some specialized knowledge could be the underlying reasons (Wolfe & Smith, 2007). However, our result is in agreement with that of Beglar's (2010) study, showing that the response processes that test takers had were in line with those that item writers had had in mind, with those provoked by test items.

As for consequential validity, the evidence seems to be inconclusive. The result seems not to be consistent with that of Baghaee and Amrahi's (2011) study. A possible explanation might be the large number of items used in that study and another plausible explanation may be that all levels of ability seemed to have been targeted. Furthermore, our finding appears to be in partial agreement with that of some research study (e.g., Ravand & Firoozi, 2016). The inspection of item infit and outfit in the current study revealed misfitting items and the person-item map showed pronounced gaps. The former finding is in contrast to that of Ravand and Firoozi's (2016) research study, and the latter result echoes theirs. As the test items of our study seem not to fully cover all person ability levels and some items were misfitting, the evidence to endorse this aspect of validity appears to be less definitive.

The current study has some contributions to the Iranian context wherein the test is administered. Scant research on the validity of this test and the lack of study on its recent versions, including the listening part, urge us to examine the interpretations and uses of test scores which have severe consequences for examinees. The results have implications for test developers and psychometricians alike to shed light on how the measure works and assist them in test score interpretations and uses when it comes to item designing. For instance, test developers and item writers are advised to consider targeting a wide range of ability levels with full coverage and ensuring content validity. Furthermore, it seems that the item specification of this version of the UTEPT test may target redundant language features and should be respecified. To distinguish between different levels of ability, all items should not be too challenging or too easy such that all examinees either fail or endorse them, resulting in no variation in their scores. Otherwise, the function of the item(s) will be called into question, and the external aspect of the validity will not be espoused. Because its content, external, and consequential aspects of validity were under question, the validation process should be repeated to make some revisions to the test such as designing less difficult items and covering a wide range of question types from basic to inferencing questions in the reading section. As for the listening section, the number of items should be increased to cover the construct more comprehensively and increase reliability. Test developers and item writers alike should also consider less challenging listening items by making use of a combination of conversations and lectures because conversations have a shorter duration and in comparison with lectures including scientific issues seem to include more general content. These can be considered in future versions of the test. As validation is a process, high-stakes test developers, in general, and UTEPT test designers, in particular, should make use of the Rasch model and Messick's framework continually to check for their test validity as disregarding each aspect may jeopardize the validity of the test interpretation and use.

Limitations and future directions

The current study strived to investigate the validity of the new version of the UTEPT test; however, the findings were subject to some limitations. One limitation is related to probing generalizability focusing on gender, which was due to the fact that no further demographic information regarding the testees was available. Hence, future studies can make use of other examinee-related background variables such as academic majors to examine the generalizability aspect of validity. Another limitation is concerned with not having access to the testees to investigate possible alignment between perceived difficulty and estimated difficulty. To throw light on the difficulty of items, qualitative and quantitative explorations, along with the use of cognitive psychology measures, are recommended. Cognitive load measures and difficulty estimates of the Rasch model can paint a more comprehensive picture of item difficulty and its functioning (Noroozi & Karami, 2022). Also, the think-aloud technique can unravel the processing of test items, which is used for checking the construct validity (Ary et al., 2019). Future studies can also delve into the reasons why the items (e.g., listening section) cannot make a distinction between the examinees, whether the reasoning may lie in the small number of items, leading to incomprehensive coverage of wide ability levels, or in the difficulty of items.

Abbreviations

DIF	Differential item functioning
EFA	Exploratory factor analysis
GE	General English
IUEE	Iranian University Entrance Examination
LR	Logistic regression
MTMM	Multi-trait multi-method
PCA	Principal component analysis
UTEPT	University of Tehran English Language Proficiency Test

Acknowledgements

We would like to thank the editor and reviewers.

Authors' contributions

The first author did the Rasch analysis and wrote and edited the paper. The second author provided the data for the analysis and read and edited the paper. The authors read and approved the final manuscript.

Authors' information

Shadi Noroozi (Shadi.Noroozi@ut.ac.ir) is a PhD candidate in Applied Linguistics/TESOL at the University of Tehran, Iran. She teaches general English to undergraduates at the University of Tehran, Iran. Her research interest is Educational measurement, validity, and the Rasch model in the context of language testing. Also, her areas of interest encompass teaching academic reading and writing.

Hossein Karami (hkarami@ut.ac.ir) is an associate professor of Applied Linguistics/TESOL at the English Department of the University of Tehran, Iran. His areas of interest include validity and fairness, especially in the context of language testing. His research has been published in various international scholarly journals including ***Language Testing***, ***International***

Journal of Bilingual Education and Bilingualism, *Educational Research and Evaluation*, *RELC Journal*, *Psychological Test and Assessment Modeling*, *TESOL Journal*, *Asia-Pacific Education Review*, and *International Journal of Language Studies*.

Funding
We received no funding for this research.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 3 January 2024 Accepted: 9 May 2024
Published online: 20 May 2024

References

Alavi, S. M., & Bordbar, S. (2020). Detecting gender-biased items in a high-stakes language proficiency test: using Rasch model measurement. *International Journal of Quantitative Research in Education, 5*(3), 227–310. https://doi.org/10.1504/IJQRE.2021.119817

Alavi, S. M., Kaivanpanah, S., & Nayernia, A. (2011). The factor structure of a written English proficiency test: A structural equation modeling approach. *Iranian Journal of Applied Language Studies, 3*(2), 27–50. https://doi.org/10.22111/ijals.2011.1008

Alavi, S. M., Rezaee, A. A., & Amirian, S. M. R. (2012). Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning, 3*(7), 39–65. Amirian, S.M.R., Alavi, S.M., & Fidalgo, A.M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, *4*(2), 187-203.

Amirian, S.M.R., Alavi, S.M., & Fidalgo, A.M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing, 4*(2), 187–203.

Amirian, S. M. R., Ghonsooly, B., & Amirian, S. K. (2020). Investigating fairness of reading comprehension section of INUEE: learner's attitudes towards DIF sources. *International Journal of Language Testing, 10*(2), 88–100.

Ary, D., Jacobs, L. C., Irvine, S., & Walker, D. (2019). *Introduction to research in education* (10th ed.). Boston, MA: Wadsworth Cengage Learning

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing, 40*(1), 8–14. https://doi.org/10.1177/02655322221125204

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford 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. https://doi.org/10.4304/jltr.2.5.1052-106

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: L. Erlbaum

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York: Routledge

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage

Engelhard, G., & Wind, S.A. (2017). *Invariant measurement with raters and rating scales: Rasch models for ratermediated assessments* (1st ed.). New York: Routledge

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing, 31*(4), 433–451. https://doi.org/10.1177/0265532214526748

Gu, L., Lockwood, J., & Powers, D. E. (2015). Evaluating the TOEFL Junior® standard test as a measure of progress for young English language learners (Research Report No. RR-15–22). Educational Testing Service. https://doi.org/10.1002/ets2.12064

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the listening vocabulary levels test. *Language Testing in Asia, 11*(1), 16. https://doi.org/10.1186/s40468-021-00132-7

Heydari, P., Bagheri, M. S., Zamanian, M., Sadighi, F., & Yarmohammadi, L. (2014). Investigating the construct validity of structure and written expression section of TOLIMO through IRT. *International Journal of Language Learning and Applied Linguistics World, 5*, 115–123.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 167–178.

Karami, H. (2013). An investigation of the gender differential performance on a high-stakes language proficiency test in Iran. *Asia Pacific Education Review, 14*(3), 435–444. https://doi.org/10.1007/s12564-013-9272-y

Karami, H., & Khodi, A. (2021). Differential item functioning and test performance: A comparison between the Rasch model, logistic regression and Mantel-Haenszel. *Journal of Foreign Language Research, 10*(4), 842–853. https://doi.org/10.22059/jflr.2021.315079.783

Khodi, A., Alavi, S. M., & Karami, H. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Language Testing in Asia, 11*(14), 1–10. https://doi.org/10.1186/s40468-021-00125-6

Khodi, A., Ponniah, L. S., Farrokhi, A. H., & Sadeghi, F. (2024). Test review of Iranian English language proficiency test: MSRT test. *Language Testing in Asia, 14*(4), 1–11. https://doi.org/10.1186/s40468-023-00270-0

Linacre, J. M. (2013). Winsteps® (version 3.80.1) [Computer Software]. Winsteps.com.

Linacre, J. M. (2021). *Winsteps® Rasch measurement computer program user's guide*. Winsteps.com.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266–283.

Liu, T., Aryadoust, V., & Foo, S. (2022). Examining the factor structure and its replicability across multiple listening test forms: Validity evidence for the Michigan English Test. *Language Testing, 39*(1), 142–171. https://doi.org/10.1177/02655322211018139

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741–760. https://doi.org/10.1177/1362168814567889

McNamara, T. F. & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: American Council on education and Macmillan

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Michell, J. (2004). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, England: Cambridge University Press

Motallebzadeh, K., & Khosravani, M. (2020). Comparing predictive validity of Islamic Azad University English Proficiency Test and standard proficiency tests against a socio-cognitively validated test of English for specific purpose. *International Journal of Linguistics, Literature and Translation, 3*(12), 68–72. https://doi.org/10.32996/ijllt.2020.3.12.9

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia, 12*(1), 1–19. https://doi.org/10.1186/s40468-022-00163-8

Priest, G. (2000). *Logic: A very short introduction*. Malden, MA & Oxford: Oxford University Press

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203–219. https://doi.org/10.1177/02655322231202947

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing, 6*(1), 1–23.

Rezaee, A. A., & Salehi, M. (2009). The construct validity of a language proficiency test: A multitrait multimethod approach. *Teaching English Language, 3*(1), 93–110. https://doi.org/10.22132/tel.2009.128679

Rezaee, A. A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test. *Research in Contemporary World Literature, 14*(56), 89–108.

Salehi, M. (2011). On the factor structure of a reading comprehension test. *English Language Teaching, 4*(2), 242–249.

Salehi, M. (2012). The construct validity of a test: A triangulation of approaches. *Language Testing in Asia, 2*(2), 102–119. https://doi.org/10.1186/2229-0443-2-2-102

Salehi, M., & Tayebi, A. (2012). Differential item functioning (DIF) in terms of gender in the reading comprehension subtest of a high-stakes test. *Iranian Journal of Applied Language Studies, 4*(1), 135–168. https://doi.org/10.22111/ijals.2012.1351

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317. https://doi.org/10.1177/026553229601300305

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (TOEFL iBT Research Report 07). *Educational Testing Service*. https://doi.org/10.1002/j.2333-8504.2008.tb02152.x

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

Wright, B., D., & Stone, M. H. (1979). *Best test design. Rasch Measurement*. Chicago, IL: ERIC

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: Mesa Press

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

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Publisher's Note