Journal of English language Teaching and Learning

University of Tabriz

Volume 14, Issue 29, (Spring & Summer 2022)

Applying a two-parameter item response model to explore the psychometric properties: The case of the ministry of Science, Research and Technology (MSRT) high-stakes English Language Proficiency test

Shahram Ghahraki

English Language and Literature Department, University of Isfahan, Isfahan, Iran

sghahraki@fgn.ui.ac.ir

Mansoor Tavakoli (Corresponding Author)

Applied Linguistics Department, University of Isfahan, Isfahan, Iran

tavakoli@fgn.ui.ac.ir

Saeed Ketabi

Applied Linguistics Department, University of Isfahan, Isfahan, Iran

ketabi@fgn.ui.ac.ir

ARTICLE INFO:

Received date:
2021.05.31
Accepted date:
2021.12.28

Print ISSN: 2251-7995
Online ISSN: 2676-6876

Keywords:

IRT, MSRT, High-stakes, Item analysis, Item difficulty, Item discrimination, Accountability

Abstract

Perhaps the degree of test difficulty is one of the most significant characteristics of a test. However, no empirical research on the difficulty of the MSRT test has been carried out. The current study attempts to fill the gap by utilizing a two-parameter item response model to investigate the psychometric properties (item difficulty and item discrimination) of the MSRT test. The Test Information Function (TIF) was also figured out to estimate how well the test at what range of ability distinguishes respondents. To this end, 328 graduate students (39.9% men and 60.1% women) were selected randomly from three universities in Isfahan. A version of MSRT English proficiency test was administered to the participants. The results supported the unidimensionality of the components of MSRT test. Analysis of difficulty and discrimination indices of the total test revealed that 14% of the test items were either easy / very easy, 38% were medium, and 48% were either difficult or very difficult. In addition, 14% of the total items were classified as nonfunctioning. They discriminated negatively or did not discriminate at all. 7% of the total items discriminated poorly, 17% discriminated moderately, and 62% discriminated either highly or perfectly, however they differentiated between high-ability and higher-ability test takers. Thus, 38% of the items displayed satisfactory difficulty. Too easy (14%) and too difficult (48%) items could be one potential reason why some items have low discriminating power. An auxiliary inspection of items by the MSRT test developers is indispensable.

DOI: 10.22034/ELT.2021.46325.2396

Citation: Ghahraki, S., Tavakoli, M., Ketabi, S. (2022). Applying a two-parameter item response model to explore the psychometric properties: The case of the ministry of Science, Research and Technology (MSRT) high-stakes English Language Proficiency test. *Journal of English Language Teaching and Learning*, 14(29), 1-26. Doi: 10.22034/ELT.2021.46325.2396

1. Introduction

English language tests are widely used for screening participants in Iranian university entrance examinations. The Iranian universities, which offer PhD programs, only accept those applicants who prove to have a reasonable degree of English proficiency which is usually measured either by an English test conducted by the same university (eg., University of Tehran's English Proficiency Test, The English Examination of Tarbiat Modares University, University of Isfahan's English Proficiency Test) or by a test like the Ministry of Science, Research, and Technology English proficiency test (MSRT). All PhD students have to provide proof of language proficiency before their comprehensive examinations. Each year more than 10000 PhD applicants who are graduates of different fields from different universities participate in an MSRT nationwide high-stake exam (Sahraee & Mameghani, 2013). The MSRT is similar to the TOEFL-PBT in its structure, and it examines the candidates' ability to listening comprehension, structure and written expression, reading comprehension and vocabulary. Admission to this TOEFL like test is a prerequisite for those who are going to pursue their prospective PhD subject matter at the state universities.

1.1 MSRT English Proficiency Test

MSRT is an abbreviation of Ministry of Science, Research, and Technology, formerly known as Ministry of Culture and Higher Education (MCHE). Geranpayeh (1994) reported the first version of the test was designed as a prerequisite exam for screening Iranian graduate students who were awarded a scholarship to continue their studies towards a PhD degree in English speaking countries before sitting TOEFL or IELTS. The test has been administered since 1989.

≠ *Test purpose*: The MSRT test is designed to assess the English language proficiency of applicants who plan to study in PhD program (www.iranscholarship.msrt.ir) in Iran.

≠ *Test use*: The MSRT test is used for admission into state universities in Iran (www. iranscholarship.msrt.ir; www.rasanews.ir).

≠ *Registration*: Test takers must register online. After the successful registration, test takers will receive a confirmation registration along with login password for the follow up the test results.

≠ *Price*: The test registration fee is the same for all applicants. Detailed information is available at the MSRT official website.

≠ *Administration*: The test is administered seven to nine times a year at the test centers in many large universities in Iran.

≠ *Test length*: The test consists of three sections, each separately timed: Listening Comprehension (30 items, 35 minutes), Structure and Written Expression (30 items, 20 minutes), and Reading Comprehension and Vocabulary (40 items, 45 minutes).

≠ *Scores and scoring procedures*: The questions are weighted equally. For each correct response, one point is considered. There is no penalty for incorrect response. The scores of all the three sections are reported within seven working days after the test administration. The total score is computed on the basis of the sum of the three sections (0-100). Overall score reports are viewed via a test taker's online account at MSRT official website. If

applicants believe that their test score results are incorrect, they can send a re-score request for further consideration. Test takers who gain the total score over 50, the minimum criterion (passing score) will receive an official certificate which is valid for two years from the test date (Farhady & Hedayati, 2009). Those who fail the test can take the test multiple times.

≠ *Author and publisher*: The Ministry of Science and Research Technology.

≠ *Contact information*: See https://saorg.ir.

Salehi (2011) investigated the construct validity of the reading section of the University of Tehran English Proficiency Test (UTEPT). In another study, Kiani and Haghighi (2006) examined the reliability issues of the English Examination of Tarbiat Modares University (TMU). They reported the test was too difficult for the candidates. A number of candidates' complaints have been filed against the MSRT test. They can be found at www.phdazmoon.net. Some of them were as follows: "I have a problem with this (MSRT) test", "the test have not been running smoothly", "if it goes on like this we will lose the PhD program", "according to the number of test questions, the time was not enough and I couldn't answer all the questions", "the listening and reading sections were difficult", "the reading section was very long and I couldn't finish it". The test information utilized within MCHE/MSRT is practically never revealed, apparently for security reasons. One of the main threats to the construct validity is construct-irrelevant factors (Messick, 1989). Haladyna & Downing (2004) identified various construct-irrelevant sources that widely threaten test scores interpretations in high-stakes tests. The above complaints may be identified as construct-irrelevant variance.

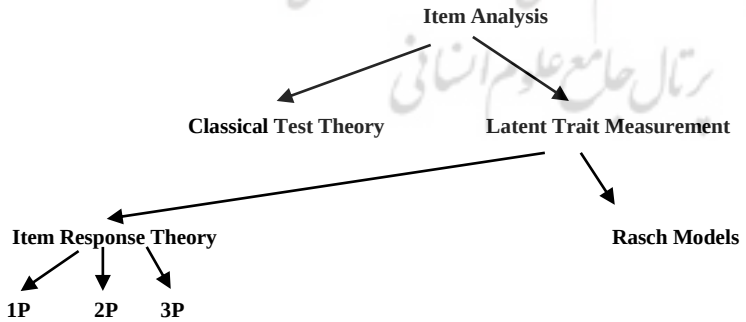
The trouble of access to the test data and the absence of test design documentation and implementation cast doubt on the test that it might be biased. Farhady and Hedayati (2009) demonstrate that no written report is available on the psychometric aspects of tests made by the MSRT. Geranpayeh (1994, p.57) points out, "There are no published data about the validity and reliability of this (MCHE) test". In the same vein, Farhady and Hedayati (2009) acknowledge that no third party has been allowed to scrutinize the test. Few studies have been carried out (Fallahian & Tabatabaei, 2015; Sahraee & Mameghani, 2013; Noori & Hosseini Zadeh, 2017) on the MSRT. Using a quantitative approach, Fallahian & Tabatabaei (2015) investigated the construct validity of reading section of the MSRT test. They questioned the validity of the MSRT reading module and reported the items did not assess the reading skills. Sahraee and Mameghani (2013) studied the reliability and validity of the MSRT. Taking Pearson product-moment correlation coefficient and comparing the means of ten versions (2010/7/8-2011/9/9) of the MSRT tests, they claimed that in general the whole test had adequate validity but it did not have the necessary validity at its components. Validity conceptualization has been changed since Messick's seminal paper in 1989. Validity is no longer is the property of a test, it refers to use of a test for specific purpose (e.g., American Psychological Association, American Educational Research Associations, & National Council on Measurement in Education, 1991; Bachman & palmer, 2010; kane, 1992; Messick, 1989; Shepard, 1993). Discussions about the different definitions of validity are beyond the scope of this article. Noori & Hosseini Zadeh (2017) reviewed the merits and demerits of the MSRT test descriptively.

It seems that no attention has been paid to the difficulty levels of the test. Perhaps one of the most significant characteristics of an item is its difficulty. Henning (1987:49) pointed out, "…when tests are rejected as unreliable measures for a given sample of examinees, it is due not so much to the carelessness of the item writers as to the misfit of item difficulty to person ability". Items that are too difficult or too easy for a given test takers influence the reliability and validity of a test. In other words, the more the quality of the items on a test are improved, the more the overall quality of the test improves-hence improving both reliability and validity. The degree of difficulty and the ability to discriminate are two fundamental considerations in test quality and they can be determined by item analysis (e.g., Airasian, 1988; Bachman, 2004; Bachman & Eignor, 1997; Baker, 1977, 1989; Boopathiraj & Chellamani,2013; Camilli & Shepard, 1994; Cohen, 1980; Deville & Chalhoub-Deville, 1993; Downing & Haladyna, 2006; Farhady, Jafarpoor, & Birjandi, 1994; Gilbert & Newtton, 1997; Green, 2013; Haladyna, 2004, 2016; Haladyna, & Rodriguez,2013; Malec, & Krzemińska-Adamek, 2020; Moss, 2017; Osterlind, 1998; Wiersma, & Jurs, 1990; Wright, 2008).

1.2. Item analysis

Item analysis is a way of measuring the quality of a test by investigating how suitable the test items were for the test takers and how well they measured their ability. It is used to identify the items which are too difficult or too easy. It enables discriminating between good and weak students. It also provides a way of re-using items over and over again in different tests (e.g.,Green, 2013,2019; Haladyna, 2016; Moss, 2017; Mousavi, 2009). There are two analytical approaches employed in item analysis to analyze the items: classical test theory (CTT) and latent trait models (e.g., Henning, 1987; Bachman, 2004). As Figure 1.1 shows items can be analyzed through classical item analysis which is based on classical test theory (Bachman 2004), and latent trait measurement theory or item response theory (IRT) (Henning 1987). The statistical procedures in item response theory includes a one-parameter model (Rasch model); a two-parameter IRT model; and a three- parameter IRT model. Each will be discussed briefly below.

Figure 1.1. The family of the analytical tools used in item analysis in analyzing the items



1.2.1 Classical test theory

The emergence of CTT dates back to early 20th century with a major influence on the course of measurement and reliability concept (e.g., Brown, 2012; Traub, 1997). Despite the limitations of CTT, many of the principles and techniques emerged from this approach are still widely used today (Brown, 2013). It has been argued that CTT model is based on a number of assumptions. The basic premise underlying CTT is that the observed score consists of two

elements: a true score that is the true ability of the test taker, and an error score that is due to factors other than the ability intended to measure. It has also assumed that the error is normally distributed, uncorrelated with true score, and has a mean of zero. CTT is concerned with the relationship between the true score and error score and because measurement devices are subject to errors, the score one gets, cannot be a true manifestation of his/her ability. The observed score is, therefore, an unreal score because that particular observed score has an error in it and the greater the error the smaller is the true score. The CTT model also assumes that all errors are unsystematic, random and uncorrelated with true score. They are unsystematic because they are unpredictable and they are unrelated to the true score. CTT treats error variance as homogeneous in origin; therefore, CTT fails to distinguish random errors from systematic errors, and it is defined as the variance of true scores. It does not really identify the multiple sources of variance and how they interact (e.g., Bachman, 2004; Kline, 2005).

According to several canonical books on testing and measurement (e.g., Bachman, 2004; Brown, 2005; Crocker & Algina, 1986) CTT provides measures and statistics both at the test level (reliability) and at the item level (item difficulty and item discrimination). On the basis of the relationship between the true score variance and error score variance, CTT focuses on a variety of reliability formula (e.g., Cronbach's alpha, KR20, KR21, split-half reliability) for measuring the consistency of assessment instruments (e.g., Brown, 2013; Sawaki, 2013). Two prominent measures of item analysis are item facility (IF) and item discrimination (ID). IF is also referred to as "item difficulty" which is the most commonly used term and labeled as p-value. Item difficulty is the proportion of correct responses for every single item. It reflects easiness of the item for a specific group of participants. Item difficulty index ranges from 0-100% or 0.0-1.0. Items with facility indexes beyond 0.70 are relatively easy, and items with facility indexes below 0.30 are relatively difficult. In a proficiency test, item facility values that fall around 50% are said to be optimal (Popham, 2000). In general, items with facility indexes within the range of 0.30-0.70 or 30%-70% are often quoted (Bachman, 2004; Brown, 2005) as being recommended or acceptable in language proficiency tests, yet items with facility indexes "between 20% and 80% can also be useful (see Green, 2013) provided the items discriminate and contribute to the internal consistency of the task" (Green, 2019:23). As a rule of thumb, items with p-values of less than 0.33 or greater than 0.67 are considered to be misfitting and should be rejected (Henning, 1987; Reynolds, Perkins, and Brutten 1994). This reference range is not necessarily absolute. Farhady et al. (1994), and Kiany & Haghighi (2006) advocate rejection of items outside of the range of 0.37 to 0.63.

The other test item statistic is item discrimination (ID). Item discrimination refers to how well a test item discriminates between weak and strong examinees in the ability being tested. It is an index deriving from comparing the difference between the performance of high achieving and the low achieving examinees (Farhady et al., 1994). Therefore, ID is the degree to which an item discriminates the more knowledgeable test-takers from the less proficient test-takers. The higher the value of ID, the more the item could separate between test- takers of higher and lower abilities. In general, the range of the discrimination index is -1.0 to 1.0; nevertheless, "items which show discrimination value beyond 0.40 can be considered acceptable" (Farhady et. al, 1994:104).

1.2.1.1 Limitations of classical item analysis

Although CTT has been the psychometric backbone of achievement testing and has served many monumental contributions to the measurement community for most of this century, certain limitations remain (e.g., Henning, 1984; Kohli, Koran, and Henn, 2015). Bachman (2004) points out five deficiencies of CTT, but here we are mainly interested in a very important limitation, "classical item analysis are essentially sample-based descriptive statistics" (Bachman, 2004:139). This means that in the CTT framework, the person and item statistics are test- and sample-dependent. Specifically, item statistics (item difficulty and item discrimination) are dependent on the sample of examines selected to answer the items. In other words, if one is going to administer that test into a different group of respondents even if the same number of items or types of items are going to be used, different results might be achieved. Likewise, the scores received by respondents depend on the collection of items they have been asked to answer. Therefore, making generalizations across different groups of respondents or across different test formats may not be possible (Bachman, 2004; Henning, 1984; Kohli et.al, 2015; Janssen, Meier, and Trace, 2014). Fan (1998) labeled this limitation as "circular dependency" i.e. person observed scores are dependent on the item statistics (i.e., item difficulty and item discrimination) and item statistics is dependent on observed scores. "Thus, person true score estimates are not invariant across different item sets, and item property estimates are not invariant across different person samples. This imparts a particular difficulty in comparing true scores across different assessments" (Frey, 2018:379). In sum, classical item analysis is carried out on the test as a whole rather than on the item and although item statistics can be made, they apply only to the particular group, or sample, of respondents on the particular set, or sample, of items that make up the test. Because of the dependency of item statistics on a specific sample- from practical perspective- it is difficult for classical item analysis to deal with the more complex testing situations, such as assessing respondent performance at different points in time; administering multiple test forms which contain different items of different difficulty of the test to several different groups at the same time for security reasons (Bachman, 2004).

1.2.2 Item response theory

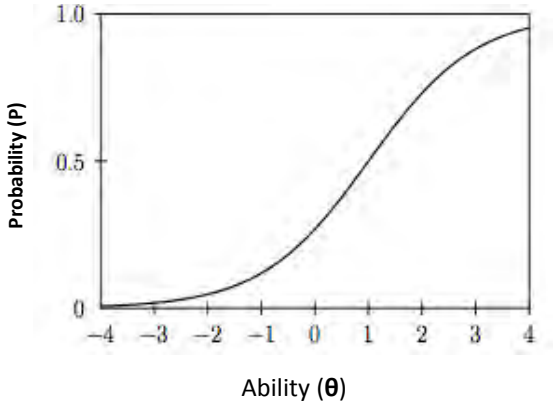
The beginning of latent trait models can be traced back to 1940s (quoted in Carlson and Davier 2013), but the models were not widely used until 1960s due to the lack of specialized software. As the name implies, these models assume to measure the underlying latent trait (or ability) which is making the test performance rather than measuring performance per se. Latent trait models (also known as item response models) and the theory upon which they are based on is called item response theory (IRT).

Unlike CTT, in which the unit of analysis is the whole test (total score or mean), in IRT, the unit of analysis is the individual item. Within the IRT framework, the main concern is that a respondent's performance on each individual item is estimated by the respondent's level of ability, as well as, the characteristics of the item. Item characteristics are independent on the specific sample and invariance is assumed across various populations and among different groups of respondents. According to Hambleton and Swaminathan (1985, quoted in Bachman 2004:142) "item parameter estimates are independent of the group of examinees used and test taker ability estimates are independent of the particular set of test items used". Therefore, IRT

provides both sample-free item calibrations and test-free person measures. This leads to them being sample-free.

According to Bachman (2004) popular IRT models are named on the basis of the number of parameters they include. If one parameter of the item is included (e.g., item difficulty) or the "b"-parameter, the IRT model is called the 1-parameter model (1PM) which is sometimes referred to as the Rasch model. Rasch model is distinct from IRT in the philosophy, origin, and history but the way it is being computed is similar to the one- parameter IRT. IRT stands towards "fitness" (fit the model to the data) and it is said to be descriptive whereas Rasch is prescriptive (fit the data into the model). Rasch inclines to "parsimony" (simplicity) because it stays with one parameter only (e.g., Andrich, 2004, 2010; Wright and Stone, 1979; Yu, 2010). If two characteristics of the item are modeled (e.g., item difficulty and item discrimination) or the "a" parameter, the IRT model is called the 2-parameter model (2PM). When three characteristics of the item are modeled (e.g., item difficulty, item discrimination, and a pseudo-chance, or guessing) or the "c" parameter, the IRT model is called the 3-parameter model (3PM). IRT is based on a number of assumptions, the first one being unidimensionality assumption. It "implies that a test should measure one single construct or dimension at a time (Baghaei Moghadam, 2009:19). This assumption cannot be completely met because of different factors such as the test anxiety, motivation, test taking qualities, and personality-related factors, etc. Unidimensionality is relative and not an absolute matter; it is a matter of degree (e.g., Andrich, 1988; Henning, Hudson and Turner, 1985; Baghaei Moghadam, 2009). The second one is local independence. It is assumed that item responses in a test are unrelated to one another. If two items are locally independent, then success or failure on one item does not affect the probability of succeeding on the other item. When the assumption of unidimensionality holds true, local independence is obtained (Lord, 1980). A specific assumption is that each test taker responding to a test item has some amount of the underlying ability. The relationship between the test takers' underlying levels of ability and their performances on the item are represented in a mathematical formula, called item characteristic curve (ICC) or item characteristic function (ICF) (Bachman, 2004; Ockey, 2012). The ICC is the bedrock of IRT models explicitly stating the assumed relationship between a test taker's probability of getting the item correct and his/her level of ability. The curve graphically shows that as the level of trait increases, the probability of obtaining a correct response increases. A typical ICC has the general shape shown in Figure 1.2.

Figure 1.2. A typical item characteristic curve



In Figure 1.2, the horizontal axis represents the respondent's ability level which is symbolized by the Greek letter theta (θ), ranging from -4 to +4. It should be noted that although the figure above shows a range of ability from -4 to +4, the theoretical range of ability is from -∞ to +∞. Therefore, all ICCs actually become asymptotic to a probability of zero at one tail and to unity at the other tail. The limited range used in the figures is requisite to fit the curves on the computer screen reasonably and to provide a consistent frame of reference. The vertical axis represents the probability of getting the correct response, ranging from 0 to 1.

The two technical properties of an ICC addressed in this paper are item difficulty and item discrimination. Item difficulty (or b-parameter) refers to the location on the horizontal axis where the probability of getting an item right is 50% (see Figure 1.3 and 1.4 below). In other words, item difficulty is a location index and it can be found by drawing a hypothetical vertical line from the inflection point where the predicted probability equals 0.5 to the horizontal axis (ability continuum). For example, easy items function at the lower ability scale while hard items would function at the higher ability scale. Item discrimination (or a-parameter) refers to how well an item can discriminate between respondents having abilities below and above the item location (b-parameter). This characteristic represents the steepness of the ICC in its middle section. The slope of the curve is called item discrimination (or a-parameter) and can be found by getting the slope of the line tangent to the ICC at the b-parameter or the slope of the ICC where it is steepest. The steeper the curve, the better item discriminating power. The flatter the curve, the less item discriminating power. (e.g., Baker, 1985; Baker & Kim, 2017; Bulut, 2015).

Figure 1.3. Atypical item difficulty and item discrimination in ICC

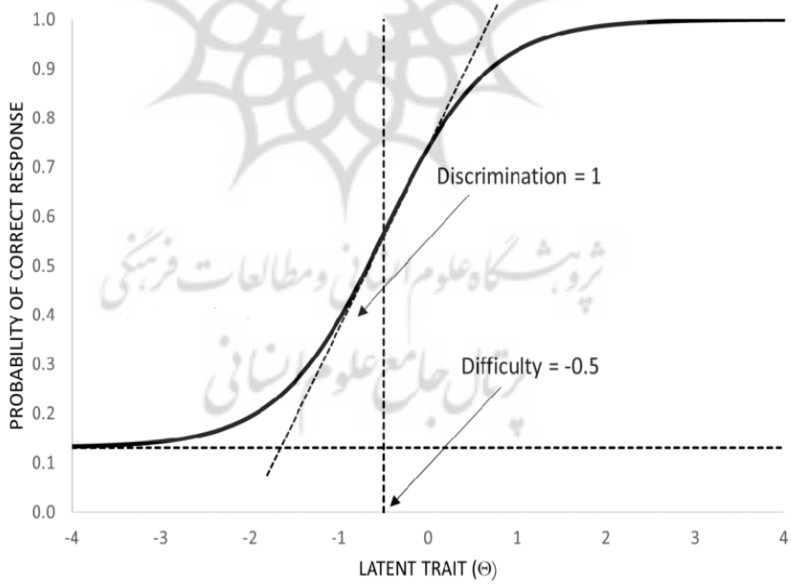
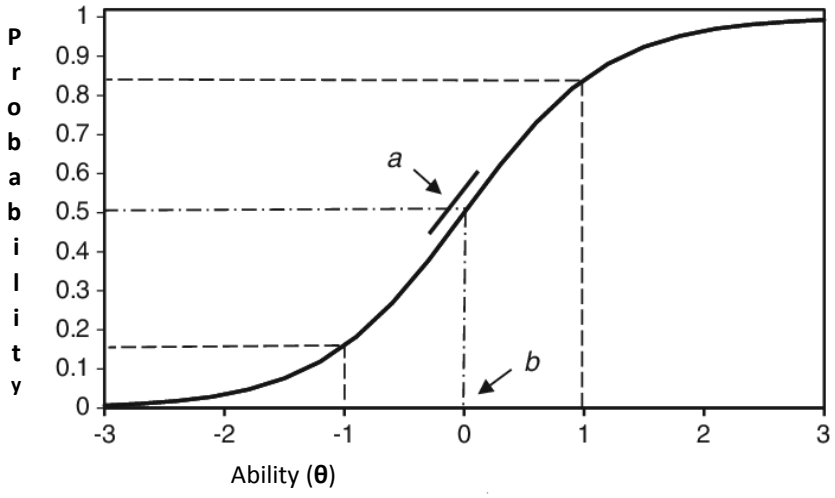


Figure 1.4. A typical a-prameter and b-parameter in ICC



Goodness of fit for IRT models

Although there are inconsistencies with the sample size requirements for assessing goodness of fit for IRT models, there are some general guidelines (Edelen & Reeve, 2007; Linacre, 1994; Morizot, Ainsworth, & Reise, 2007; Nguyen, Han, Kim, &Chan, 2014). It has been argued that for estimating simple Rasch models, the sample sizes of 100 are often adequate (Linacre, 1994). However, there are controversies over the sample sizes for more parameters of IRT models. The range of goodness of fit for more complex models have been suggested as 200 to 500 (Tsutakawa & Johnson, 1990). Linacre (1994, p.328) suggested that "at least 250 for high stakes test is enough". Similarly, Loe (2021, p.10) recommended that for "2PL (two parameter logistic model) at least 250, but best is at least 500".

As mentioned earlier, the outcome of MSRT high-stakes test is used as the sole determining factor for making a noteworthy decision; whether or not applicants should be admitted to PhD programs. This is an important issue because the higher the level of education, the more attention on the standards agenda is needed by the government. Many test takers complain about the difficulty level of test items (see1.1 above). However, to date no empirical research report is available on psychometric properties of test items made by MSRT. Consequently, the difficulty of items in MSRT test are almost never evaluated. Therefore, this issue has been considered as a significant discussion to be examined in the current study. Thus, this study aimed to answer the following research questions:

Q1. Do MSRT listening component test items have acceptable level of difficulty and discrimination power?

Q2. Do MSRT structure and written expression component test items have acceptable level of difficulty and discrimination power?

Q3. Do MSRT reading component test items have acceptable level of difficulty and discrimination power?

2. Method

2.1 Participants

The participants were 328 graduate students (39.9% men and 60.1% women) who were selected randomly from three universities in Iran, Isfahan. They were majoring either in

humanities (72%) or engineering (28%). The participants were in their last semester of university education at the level of Master's degree. They were informed by their professors and the researchers themselves that they would take part in a research, and they enthusiastically agreed to participate voluntarily. They asked the researchers to be informed about their scores in the test. Following Linacre (1994) and Loe (2021), the sample of the present study (328) seemed to meet the IRT assumptions of goodness of fit.

2.2 Instrument

The data was collected through a version of MSRT English proficiency test. The test comprised 100 items and three components, each separately timed: Listening Comprehension (30 items, 30-35 minutes), Structure and Written Expression (incomplete and incorrect structures) (30 items, 20 minutes), and Reading Comprehension and Vocabulary (40 items, 45 minutes). The total score was computed on the basis of the sum of the three sections (0-100).

2.3 Procedures

All administrative procedures of the study were followed consistently across different occasions and for all participants. The 328 participants in the study needed eight laboratory classes at various universities in Iran. This meant that 8 test sessions were carried out separately. Due to limited capacity of English laboratories (seats were available for 45 students), in each session, 41 candidates could participate in the study. Thus, eight meetings were organized for the MSRT test (8×41=328). The same procedure was used in all sessions. In every test session, necessary explanation was offered. The participants were told that the three sections of the test would be administered sequentially. The time for the test was 100 minutes.

3. Data analysis

The results obtained from test administration were analyzed through item response theory. There are various software programs for conducting IRT analysis such as IRTPRO (Cai, Thissen, & du Toit, 2011), WINSTEPS (Linacre, 2015), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002), and R-Package "ltm" (Rizopoulos, 2018), to name a few. This study used R-Package to analyze the data because of its availability and being free of charge.

4. Results

4.1 Calibration of MSRT Items through Item Response Theory

The Item Response Theory (IRT) model was run to explore the MSRT items properties; i.e. the difficulty and discrimination through a two-parameter model. The data was analyzed via the R-Package "ltm" developed by Dimitris Rizopoulos on April 18, 2018. The IRT results are discussed for three sections of MSRT followed by a discussion on test information function (TIF) and samples of item characteristic curves (ICC).

4.1.1 Testing Assumption of Unidimensionality

The characteristic of the items of the MSRT test analyzed through an IRT model. IRT assumes unidimensionality of the construct being investigated. This assumption was probed for the four components of MSRT using the method developed by Robitzsch (2019), and implemented in R-Package "sirt". Based on the results displayed in Table 4.1 it can be concluded that all four sub-sets of MSRT met the unidimensionality assumption. The values of DETECT were below .20 for all tests; except for the Incomplete section of structure which was slightly higher than

.20. The ASSI indices were all below .25 and the value of Ratios were all lower than .36. All the results supported the unidimensionality of the components of MSRT.

Table 4.1
Tests of Unidimensionality of MSRT

	Incomplete	Incorrect	Listening Comprehension	Reading Comprehension
DETECT	**0.21**	**-0.12**	**-0.12**	**-0.06**
ASSI	**0.14**	**0.01**	**-0.04**	**-0.04**
RATIO	**0.09**	**-0.06**	**-0.06**	**-0.03**

4.1.2 Listening Component of MSRT

Table 4.2 displays the item difficulty (b) and item discrimination (a) of 30 items of the listening component of MSRT. Before discussing the results, it should be noted that all tables reported in this section includes six columns as follows;

1: Items
2: Item Difficulty (b)
3: Item Discrimination (a)
4: Evaluation criteria for (b) as suggested by Baker and Kim (2017:11); i.e.

< - 2.625	Very Easy
-2.624 to -1.5	Easy
-1.49 to 0	Medium
.01 to 1.5	Hard
1.51 to 2.625	Very Hard

5: Evaluation criteria for (a) as suggested by Baker and Kim (2017:11); i.e.

-999 to 2.625	None
.01 to .4	Low
.41 to 1	Moderate
1.1 to 2.1	High
2.2 to 999	Perfect

Based on the results displayed in Table 4.2 it can be concluded that among the 30 items of the listening component of MSRT; five items were very easy, 10 were of moderate difficulty and 15 were hard.

Table 4.2
Item Difficulty and Item Discrimination of Listening Component of MSRT

Item	B	a	b Criteria	a Criteria (BK)
LM1	-3.289	0.065	Very Easy	Low
LM2	-0.279	4.941	Medium	Perfect
LM3	-0.084	1.253	Medium	High
LM4	0.156	2.453	Hard	Perfect
LM5	0.255	1.406	Hard	High
LM6	0.120	2.619	Hard	Perfect
LM7	0.691	2.121	Hard	Perfect
LM8	-6.641	-0.174	Very Easy	None
LM9	1.492	0.446	Hard	Moderate

LM10	-0.235	2.105	Medium	Perfect
LM11	-0.010	1.947	Medium	High
LM12	-0.122	1.209	Medium	High
LM13	0.219	4.106	Hard	Perfect
LM14	0.227	1.595	Hard	High
LM15	-0.095	1.596	Medium	High
LM16	1.300	1.115	Hard	High
LM17	-0.715	0.660	Medium	Moderate
LM18	-0.247	-0.402	Medium	None
LM19	-0.002	33.211	Medium	Perfect
LM20	-3.228	-0.527	Very Easy	None
LM21	0.016	2.298	Hard	Perfect
LM22	-15.237	-0.072	Very Easy	None
LM23	0.513	2.166	Hard	Perfect
LM24	0.665	1.325	Hard	High
LM25	0.026	36.805	Hard	Perfect
LM26	0.518	1.956	Hard	High
LM27	-0.595	1.022	Medium	High
LM28	-43.798	-0.030	Very Easy	None
LM29	1.004	1.358	Hard	High
LM30	0.669	1.587	Hard	High

The evaluation of the discrimination indices based on the Baker and Kim (2017) criteria indicated that five items had no discrimination, one item had low and two showed moderate discrimination. The discrimination of 12 items were high and another ten items showed perfect discriminations.

Figure 4.1 shows the Test Information Function of the listening component of MSRT. If a hypothetical vertical line is drawn from the peak of the plot to the horizontal line, the intersection, which is almost zero, shows that the test rendered the highest information about participants whose listening ability was an average one. In other words, the listening test was somehow difficult.

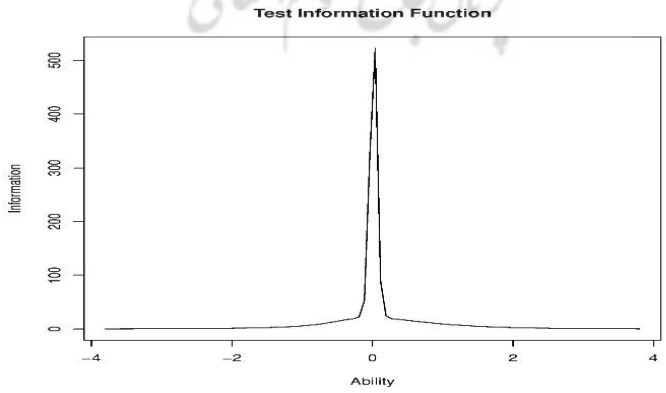


Figure 4.1 Test information function of listening component of MSRT

Figure 4.2 shows the ICC curves for the most and least difficult and discriminating items. Based on the results displayed in Table 4.2, it can be claimed that; item 28 (b = - 43.79) was

the easiest, item 9 (b = 1.492) was the most difficult, item 20 (a = - .527) and item 25 (a = 36.805) were the least and most discriminating (Figure 4.2).

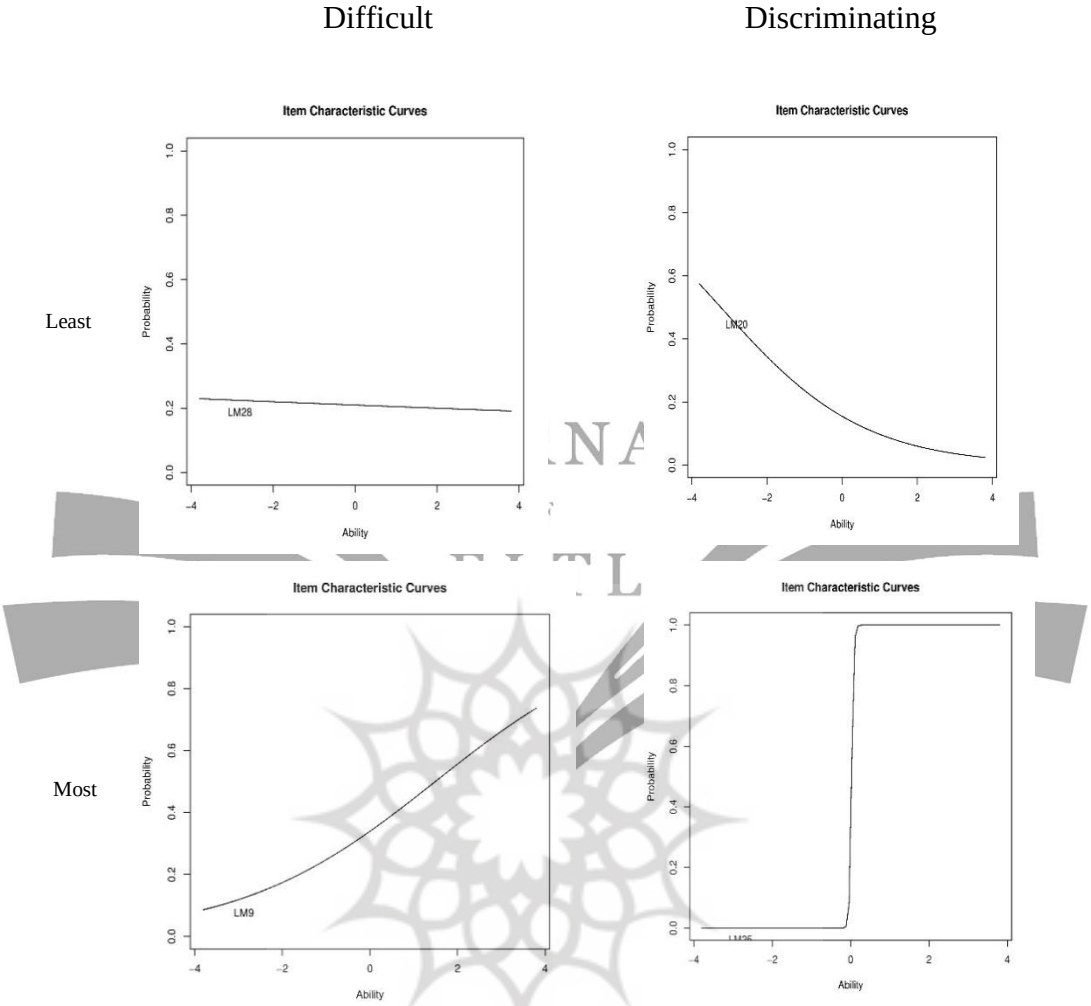


Figure 4.2 Least and most difficult and discriminating items (listening component of MSRT)

4.1.3 Incomplete Structures of MSRT

Table 4.3 displays the item difficulty (b) and item discrimination (a) of 15 items of the incomplete structures component of MSRT. Based on the results, it can be concluded that among the 15 items of the incomplete structure component of MSRT, eight items were of moderate difficulty while seven were hard.

Table 4.3
Item Difficulty and Item Discrimination of Incomplete Structures Component of MSRT

Item	B	a	b Criteria	a Criteria (BK)
SM1	-0.429	46.389	Medium	Perfect
SM2	-0.097	3.352	Medium	Perfect
SM3	-0.180	0.610	Medium	Moderate
SM4	1.005	1.286	Hard	High
SM5	-0.259	2.352	Medium	Perfect
SM6	-0.302	1.325	Medium	High
SM7	-0.351	1.937	Medium	High
SM8	0.734	2.276	Hard	Perfect

SM9	0.023	2.983	Hard	Perfect
SM10	1.140	0.989	Hard	Moderate
SM11	-0.323	0.750	Medium	Moderate
SM12	-0.175	1.284	Medium	High
SM13	0.523	0.670	Hard	Moderate
SM14	0.063	1.059	Hard	High
SM15	0.284	-0.299	Hard	None

The evaluation of the discrimination indices based on the Baker and Kim (2017) criteria indicated that one single item had no discrimination while four showed moderate discrimination. The discrimination of five items were high and another five items showed perfect discriminations.

Figure 4.3 shows the Test Information Function of the incomplete component of MSRT. If a hypothetical vertical line is drawn from the peak of the plot to the horizontal line, the intersection, which is slightly lower than zero, shows that the test rendered the highest information about participants whose knowledge on incomplete structures was below the average one.

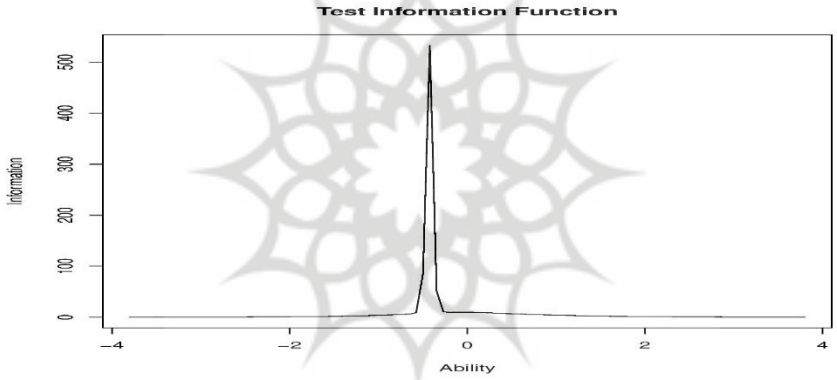


Figure 4.3 Test information function of incomplete structures component of MSRT

Figure 4.4 shows the ICC curves for the most and least difficult and discriminating items. Based on the results displayed in Table 4.3, it can be claimed that; item 1 (b = - .429) was the easiest item among the eight moderate items, item 10 (b = 1.140) was the most difficult, item 15 (a = - .299) and item 1 (a = 46.389) were the least and most discriminating items (Figure 4.4).

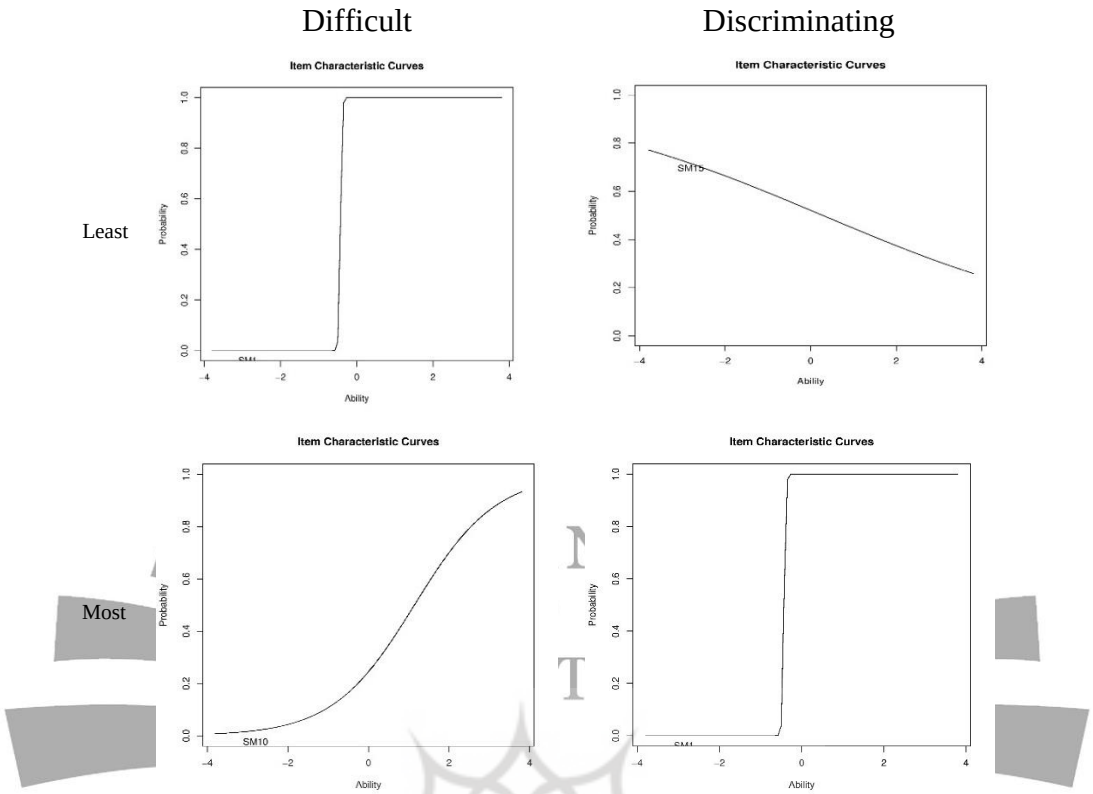


Figure 4.4 Least and most difficult and discriminating items (incomplete structures of MSRT)

4.1.4 Incorrect structures of MSRT

Table 4.4 displays the item difficulty (b) and item discrimination (a) of 15 items of the incorrect structures component of MSRT. Based on the results, it can be concluded that among the 15 items, the following difficult levels were detected: easy (one item), medium (four items), hard (eight items) and finally (two items) very hard.

Table 4.4

Item Difficulty and Item Discrimination of Incorrect Structures Component of MSRT

Item	B	a	b Criteria	a Criteria (BK)
SM16	3.675	0.436	Very Hard	Moderate
SM17	0.111	2.055	Hard	High
SM18	0.478	1.315	Hard	High
SM19	2.545	1.244	Very Hard	High
SM20	-1.171	0.167	Medium	Low
SM21	0.212	1.451	Hard	High
SM22	0.441	0.765	Hard	Moderate
SM23	-0.079	1.459	Medium	High
SM24	-2.141	-0.306	Easy	None
SM25	0.400	1.548	Hard	High
SM26	-0.395	0.758	Medium	Moderate
SM27	1.381	1.925	Hard	High
SM28	-0.379	1.132	Medium	High
SM29	0.340	1.299	Hard	High
SM30	0.152	4.442	Hard	Perfect

The evaluation of the discrimination indices based on the Baker and Kim (2017) criteria indicated that, one single item had no discrimination, one had low discrimination indices, and three showed moderate discrimination. The discrimination of nine items were high while a single item showed perfect discriminations.

Figure 4.5 shows the Test Information Function of the incorrect component of MSRT. If a hypothetical vertical line is drawn from the peak of the plot to the horizontal line, the intersection, which is slightly above zero, shows that the test rendered the highest information about participants whose knowledge on incorrect structures was slightly higher than average one.

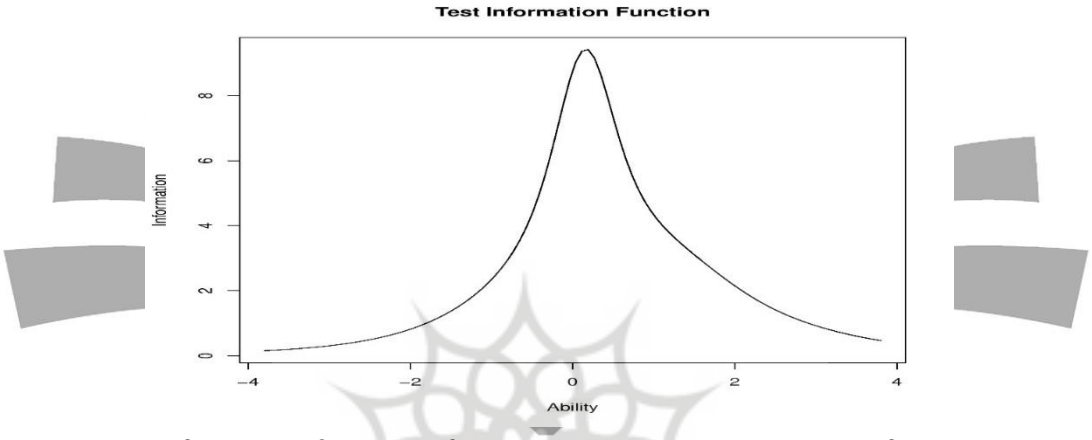
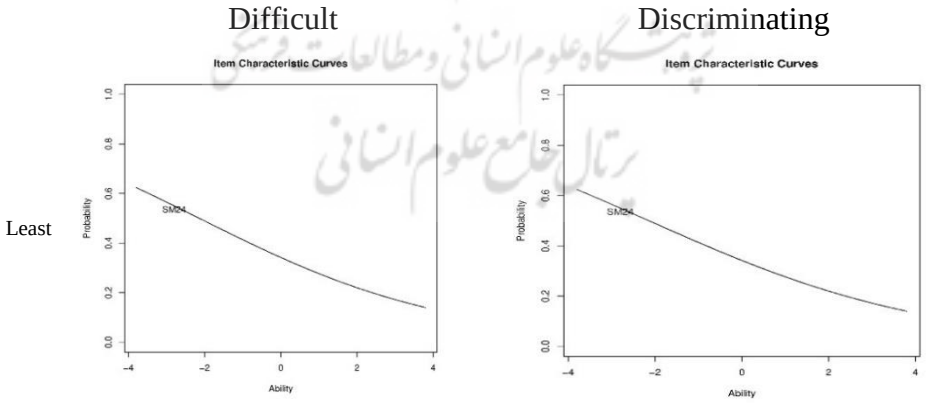


Figure 4.5 Test information function of incorrect structures component of MSRT

Figure 4.6 shows the ICC curves for the most and least difficult and discriminating items. Based on the results displayed in Table 4.4 it can be claimed that, item 24 (b = - 2.141) was the easiest, item 16 (b = 3.675) was the most difficult, item 24 (a = - .306) and item 30 (a = 4.442) were the least and most discriminating items (Figure 4.6).



Most

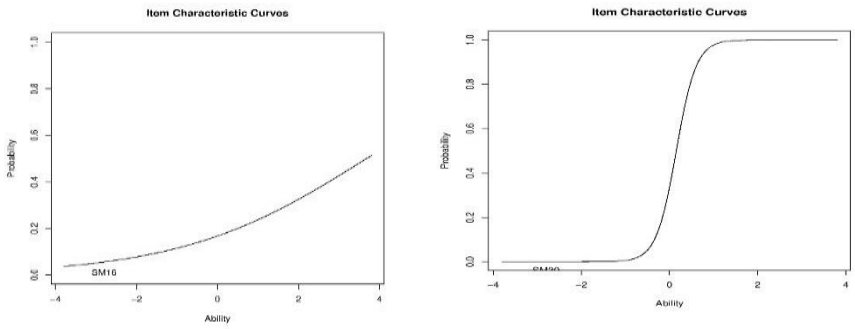


Figure 4.6 Least and most difficult and discriminating items (incorrect structures of MSRT)

4.1.5 Reading component of MSRT

Table 4.5 displays the item difficulty (b) and item discrimination (a) of 40 items of the reading component of MSRT. Based on these results it can be concluded that among the 40 items of the reading component of MSRT; eight items were either very easy or easy, 16 were of moderate difficulty and 16 items were either hard or very hard.

Table 4.5

Item Difficulty and Item Discrimination of Reading Component of MSRT

Item	B	a	b Criteria	a Criteria (BK)
RM1	-0.157	1.011	Medium	High
RM2	-0.764	2.541	Medium	Perfect
RM3	-0.357	4.013	Medium	Perfect
RM4	-0.342	0.772	Medium	Moderate
RM5	-0.242	5.400	Medium	Perfect
RM6	-0.306	4.154	Medium	Perfect
RM7	0.174	2.665	Hard	Perfect
RM8	-2.167	-0.670	Easy	None
RM9	-0.536	2.136	Medium	Perfect
RM10	-0.120	1.589	Medium	High
RM11	0.178	1.176	Hard	High
RM12	0.410	3.103	Hard	Perfect
RM13	0.413	0.774	Hard	Moderate
RM14	-0.406	1.544	Medium	High
RM15	-2.415	-0.259	Easy	None
RM16	-15.121	-0.147	Very Easy	None
RM17	-0.307	1.321	Medium	High
RM18	-2.453	-1.011	Easy	None
RM19	0.134	1.188	Hard	High
RM20	2.785	0.663	Very Hard	Moderate
RM21	-0.233	1.821	Medium	High
RM22	-0.014	0.438	Medium	Moderate
RM23	-0.135	1.376	Medium	High
RM24	0.603	1.106	Hard	High
RM25	5.240	0.539	Very Hard	Moderate
RM26	0.154	1.210	Hard	High
RM27	-2.136	-0.506	Easy	None

RM28	1.332	0.887	Hard	Moderate
RM29	2.697	0.860	Very Hard	Moderate
RM30	5.091	0.630	Very Hard	Moderate
RM31	-0.223	2.676	Medium	Perfect
RM32	-4.190	-0.320	Very Easy	None
RM33	-0.020	36.910	Medium	Perfect
RM34	1.239	0.681	Hard	Moderate
RM35	-10.127	-0.210	Very Easy	None
RM36	-0.351	2.685	Medium	Perfect
RM37	3.317	0.731	Very Hard	Moderate
RM38	-1.937	0.070	Easy	Low
RM39	0.592	0.786	Hard	Moderate
RM40	0.698	0.611	Hard	Moderate

The evaluation of the discrimination indices based on the Baker and Kim (2017) criteria indicated that, seven items had no discrimination, one item had low, and 12 items showed moderate discrimination. The discrimination of 10 items were high and another 10 items showed perfect discriminations.

Figure 4.7 shows the Test Information Function of the reading component of MSRT. If a hypothetical vertical line is drawn from the peak of the plot to the horizontal line, the intersection, which is almost zero, shows that the test rendered the highest information about participants whose reading ability was an average one. In other words, the reading test was somehow difficult.

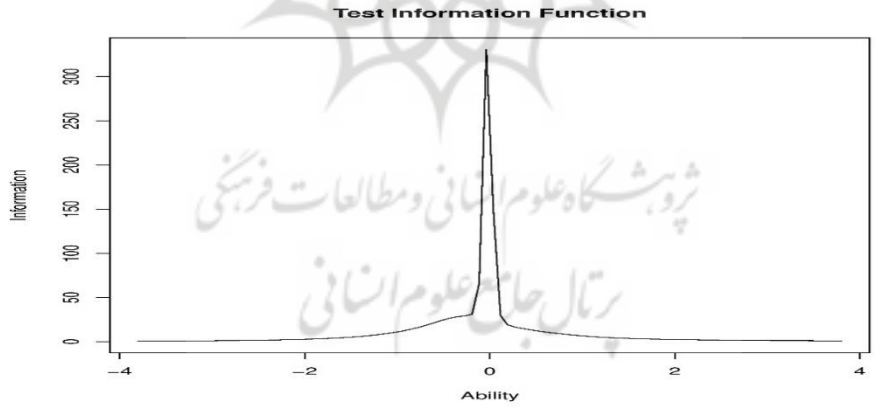


Figure 4.7 Test information function of reading component of MSRT

Figure 4.8 shows the ICC curves for the most and least difficult and discriminating items. Based on the results displayed in Table 4.5, it can be claimed that, item 16 (b = - 15.121) was the easiest, item 25 (b = 5.240) was the most difficult, item 18 (a = - 1.011) and item 33 (a = 36.910) were the least and most discriminating items (Figure 4.8).

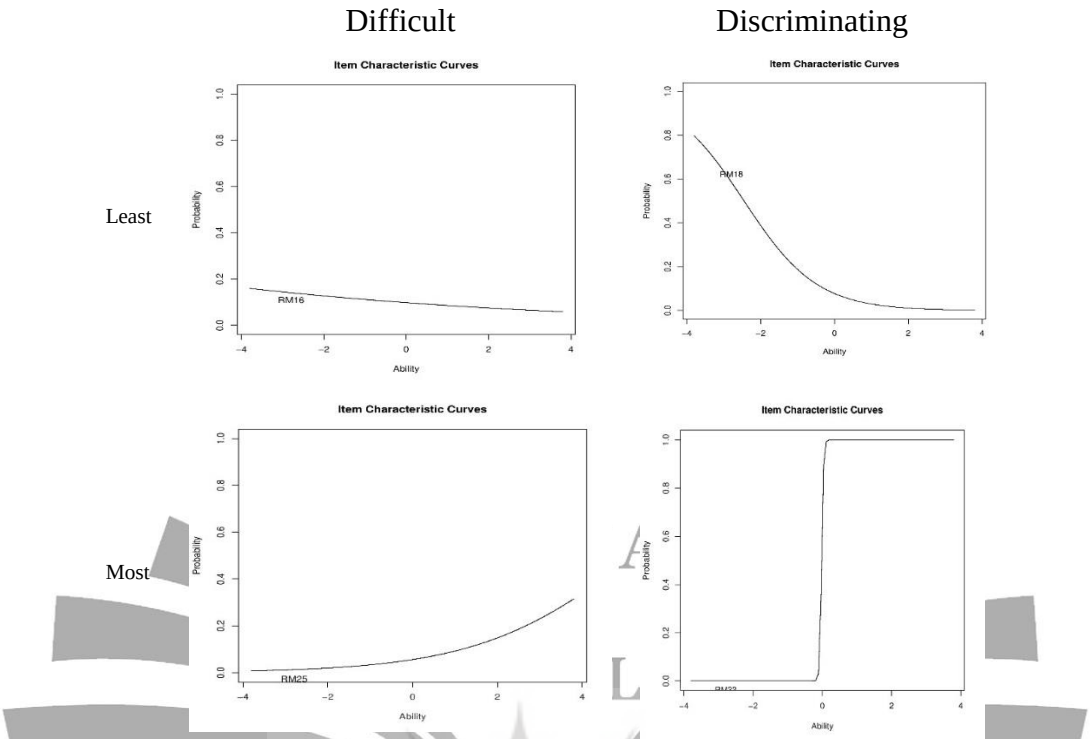


Figure 4.8 Least and most difficult and discriminating items (reading component of MSRT)

5. Discussion

The purpose of this study was to examine the psychometric properties of the MSRT test items using a 2-parameter IRT model. We assessed both item difficulty and item discrimination of the three components of MSRT test, namely listening comprehension, structure and written expression (incomplete and incorrect structures), and reading comprehension respectively. Concerning the first research question, the difficulty level and discrimination indices of listening section of MSRT test were scrutinized. The results in Table 4.2 indicated that the range of difficulty values (-43.798 to 1.492) suggested that 16.7% items were very easy, 33.3% were of moderate difficulty, and 50% were hard. The item discrimination is a useful means of adding meaning to the item difficulty interpretation. The slope of the ICC at difficulty parameter (b), indicates how well an item can discriminate between test takers in the neighborhood of this ability level (Baker & Kim, 2017). Looking down the discrimination indices in Table 4.2, the range of discrimination values (-0.030 to 36.805) indicated that 16.7% items were nonfunctioning. In other words, they were not discriminating items. These items could not differentiate between high ability and low ability test takers. In addition, items LM8, LM18, LM20, LM22, and LM28 were malfunctioning and yielded negative discrimination indices. This implies that the more knowledgeable test takers were answering the item incorrectly and the less knowledgeable test takers were answering the item correctly. Popham (2000) used a term "red flag" for this particular situation as a warning that the flawed items needed some attention. A negative discrimination index might be resulted from either the items were poorly written or there were some misleading information among the high ability test takers. Item LM1 with a value of 0.065 was low discrimination index (3.3%), while 6.6% items were moderate, 40% items were high, and 33.3% items were perfect discrimination indices. They were all functioning items. The Test Information Function (TIF) enables to estimate how

well the test at what range of ability distinguishes respondents. Looking at Figure 4.1, the TIF depicted the maximum information was at ability level of almost zero. In other words, the listening section was somehow difficult. The output in Figure 4.2 showed that the ICC for the least difficult item (LM28) was approximately linear and appeared rather flat. This was because the item difficulty value for the item was undefinable with no discrimination (Baker & Kim, 2017). The ICC for the least discriminating item (LM20) with a negative value was the opposite sign of a common assumption of IRT which the probability of correct response is positively related to the estimated ability of the test takers. The second part of Figure 4.2 graphically displayed the most difficult item (LM9). The probability of getting the item right was low for the most of the ability axis and it was increased when the higher levels were reached. The ICC for the most discriminating item (LM25) revealed a perfect discrimination. The shape of the curve was a vertical line at the x axis. To the right of the vertical line, the probability of endorsing the correct response was one and to the left of the line the probability of getting the item right was zero. Therefore, the item differentiated fully between test takers whose abilities were above and below an ability score around zero. This item was discriminating well between respondents with abilities just above and below zero.

Regarding the second research question, the results of the first half of this question (incomplete structures) in Table 4.3 showed that the range of difficulty values (-0.429 to 1.140) suggested that 53.3% items were of moderate difficulty, while 46.67% were hard. The data from Table 4.3 also indicated that the range of discrimination values (-0.299 to 46.389) implied that only one item (6.6%) was nonfunctioning, while 26.72% items were low, 33.34% items were high, and 33.34% items were perfect discrimination indices. In other words, 66.68% items were functioning and discriminating well. As shown in Figure 4.3, item difficulties were tightly clustered slightly lower than zero, therefore the TIF was peaked at this ability scale. It can be interpreted the highest information about test takers whose knowledge on incomplete structures was below average one. The ICC in Figure 4.4 indicated that item SM1 (b= -0.429) was the least difficult item. Item SM15 with a value of a= -0.299 discriminated negatively and was a malfunction item. The curve graphically illustrated that the less-ability candidates were answering the item correctly while the high-ability candidates were answering the item incorrectly. The other part of Figure 4.4 vividly showed the most difficult item (SM10) with a value of b=1.140. As the ability increases, the probability of endorsing the item correct increases. In other words, as the location of b- parameter lies toward higher ability, the more difficult the item. The curve for item SM1 with the discriminating value of a=46.389 was the most discriminating item. That is, all test takers with a theta less than 0.429 (lower-ability test takers) have an almost 0% probability of endorsing the correct answer, and all the test takers with a theta more than 0.429 (higher-ability test takers) have a 100% probability of endorsing the correct response.

The results of the second part of research question No. 2 (incorrect structures) in Table 4.4 revealed that the range of difficulty values (-2.141 to 3.675) suggested that 6.6% items were easy, 26.72% were moderate, 13.34% were hard, and 53.34% were very hard. Inspection on the results of a-parameter analysis in Table 4.4 showed that the range of item discrimination values (-0.306 to 4.442) showed that one single item (6.6%) was nonfunctioning and another one (6.6%) was low discriminating item, while 20% items were moderate, 60% were high, and

only one item (6.6%) was perfect index. The graph in Figure 4.5 shows a heavy-tailed TIF. *One reason for this could be the tails did not decrease more rapidly than an exponential ones* (Bryson, 1974*). Thus, the density of information about test takers whose ability on incorrect structures was slightly higher than the average one.* The ICC in Figure 4.6 illustrated that item SM24 (b= -2.141 and a= -0.306) was the least difficult and nonfunctioning item. It did not discriminate between high and low ability test takers. On the other hand, item SM16 with a value of b=3.675 was the most difficult. The curve shifted toward the right side of the ability scale which indicated the higher ability of the test takers got the item correctly. The curve for item SM30 with the discriminating value of a=4.442 was the most discriminating item. The item was well functioned with a pronounced slope.

With regard to the third research question, the results in Table 4.5 indicated that the range of difficulty values (-15.121 to 5.240) suggested that 7.5% items were very easy, 12.5% were easy, 40% were moderate, 27.5% hard, and 12.5% were very hard items. A closer inspection of discrimination indices in Table 4.5, reveals the range of discrimination values (-1.011 to 36.910) indicating 17.5% items were nonfunctioning, one item (2.5%) was low discriminating item, while 30% items were moderate, 25% items were high and another 25% showed perfect discrimination. Looking at Figure 4.7, it is apparent that item difficulties were tightly clustered at zero. The TIF clearly delineated that item difficulties were not widely distributed on x-axis. Looking across the ICCs in Figure 4.8, one could observe that item RM16 with a value of b=-15.121 was the least difficult and nonfunctioning item. The curve for item RM18 (a=-1.011) shifted toward the left side of the ability scale which indicated the lower ability test takers got the item correctly. The curve also graphically illustrated that the low-ability test takers were responding the item appropriately and the high-ability test takers were responding the item inappropriately. On the other hand, item RM25 with a value of b=5.240 was the most difficult. As the ability decreases, the probability of endorsing the correct answer decreases. In other words, as the location of b- parameter lies toward higher ability, the curve shifts toward the right side of the ability scale, indicating the high-ability test takers endorsed the item correctly. The curve for item RM33 with the discriminating value of a=36.910 was the strongest discriminating item. Specifically, all test takers with a theta less than 0.020 (lower-ability test takers) have an almost 0% probability of being correct or endorsing the item, and all the test takers with a theta more than 0.020 (higher-ability test takers) have a 100% probability of endorsing the correct response.

In sum, the analysis of difficulty and discrimination indices of total test revealed 14% test items were either easy or very easy, 38% were medium, and 48% were either difficult or very difficult. A basic assumption is that a test should have a widespread distribution with optimum difficulty level (68% of the test items should fall within one standard deviation of the mean) for a maximum discrimination between low-ability and high-ability test takers. In general, moderate difficulty items are preferable. In addition, 14% of the total items were classified as nonfunctioning. They discriminated negatively or did not discriminate at all. These items typically were not plausible and need to be reviewed. 7% total items discriminated poorly, 17% discriminated moderately, and 62% discriminated either highly or perfectly, however they differentiated between high-ability and higher-ability test takers. Mehrens and Lehmann (1991:167) identify one possible reason why some items have low discriminating power is "the

more difficult or easy the item, the lower its discriminating power". According to Farhady et al. (1994) too easy and too difficult items should be deleted from the test. Thus, 38% of the items displayed satisfactory difficulty, and 62% of items were not acceptable. They were either too easy (14%) or too difficult (48%). Therefore, it can be concluded that the test was difficult.

6. Conclusion

Tests are employed for a wide variety of purposes and testing is basically a decision- making endeavor. In the age of educational accountability, fair and standard tests can assist test developers to be accountable for different stakeholders. The MSRT English proficiency test is currently administered as a prerequisite for PhD comprehensive examination. The outcomes of this nationwide high-stakes test have important consequences on the final noteworthy decisions for admission to PhD programs. This is the first study to examine the item properties of MSRT test. The results of this study indicate that the test is difficult for the test takers. Additionally, the initial findings of this study disclose a number of items are either nonfunctioning or working negatively. These shortcomings may affect the range of cut- off score and test results interpretations. Therefore, MSRT test givers should be held accountable for the quality of test and some crucial caveats should be taken into account. Auxiliary inspections of items by the MSRT test developers are indispensable. Taken together, these results suggest that MSRT test authorities should establish a Standards panel to refine, revise, and redraft the test items to achieve quality assurance.

References

Airasian, P. W. (1988). *Measurement driven instruction: A closer look. Educational Measurement: Issues and Practice, 7*(4), 6-11.

Andrich, D. (1988) *Rasch Models for Measurement*. Sage Publications, Inc., Beverly Hills.

Andrich, D. (2004). Controversy and the Rash model: A characteristic of incompatible paradigm? *Medical Care, 42*(I), 1–16.

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 123– 152). New York, NY: Routledge.

American Educational Research Association, American Psychological

Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington DC.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Eignor, D. R. (1997). Recent advances in Quantitative test analysis. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, language testing and assessment, Vol. 7*, (pp. 227–242). Dordrecht: Kluwer Academic.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use*. Oxford: Oxford University Press.

Baghaei Moghadam,P. (2009). *Understanding the Rasch model*. Mashhad, Sokhangostar Publications.

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research, 47,* 151- 158.

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann.

Baker, F. B. (1989). *Computer technology in test construction and processing.* In R. L. Linn (Ed.), *Educational measurement* (pp. 409–428). Macmillan Publishing.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Berlin: Springer.

Boopathiraj, C., Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, *(2),* 189-193.Available at indianresearchjournals.com.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw- Hill.

Brown, J. D. (2012). Classical test theory. In G. Fulcher and F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp.323-335). Routledge.

Brown, J. D. (2013). Classical theory reliability. In A. Kunnan (Ed.), *Companion to language assessment, Vol. 3*. Hoboken, NJ: Wiley Blackwell.

Bryson, M. (1974). Heavy-tailed distributions: Properties and tests. Technometrics,6,61-68. http://dx.doi.org/10.1080/00401706.1974.10489150

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: Practical issues and insights. *Journal of measurement and evaluation in education and psychology, 6*(2): 313-330.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows (Computer software).*Lincolnwood, IL: Scientific Software International.

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

Carlson,J.E & D avier,M.V. (2013). *Item response theory*. Educational Testing Service, Princeton, New Jersey.

Cohen, A. D. (1980). *Testing Language Ability in the Classroom*. Rowley, Mass: Newbury House Publishers.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt, New York.

Deville, C., & Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis, and Sato's caution index used to investigate the reading recall protocol. *Language Testing, (10),* 117-132.

Downing, S. M., & *Haladyna*, T. M. (Eds.). (*2006*). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.

Edelen, M.O. & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research 16*(Suppl 1), 5-18.

Fallahian, E. & Tabatabaei, O. (2015). Construct validity of MSRT reading comprehension module in Iranian context. *English Language Teaching, 8*(9), 173-186.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/response statistics, *Educational and Psychological Measurement, 58*(3), 357- 381.

Farhady, H. & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics (29),* 132-141.

Farhady, H. Jafarpur, A. and Birjandi, P.(1994). *Language skills testing from theory to practice*. Tehran: SAMT Publications.

Frey, B. B. (Ed.). (2018).The sage encyclopedia of educational research, measurement, and evaluation. Sage Publications.

Geranpayeh, A. (1994) Are Score Comparisons across Language Proficiency Test Batteries Justified? An IELTS-TOEFL Comparability Study, *Edinburgh Working Papers in Applied Linguistics, 5:* 50-65.

Gilbert, S. & Newtton, W. J.(1997). *Principles of educational and psychological measurement and evaluation*. Wadsworth: The University of California.

Green, R. (2013). *Statistical analyses for language testers*. New York, NY: Palgrave Macmillan.

Green, R. (2019). Item analysis in language assessment. In V. Aryadoost, & M. Raquel (Eds.). *Quantitative data analysis for language assessment volume I: Fundamental techniques*. Routledge.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M. (2016). Item analysis for selected response items. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development (2nd ed)*, (pp. 392–407). New York, NY: Routledge.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: *Principles and applications.* Boston: Kluwer Academic Publishers.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing (1)*, 123–133.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research. Cambridge*: Newbury House Publishers.

Henning, G., Hudson, T. and Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing (2)*, 141–154.

Janssen, G., Meier, V., Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colombian Applied. Linguistics Journal. 16 (2),* 167–184.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, (112)*, 527-535.

Kiani,G.R. & Haghighi, M.(2006). The investigation of the TMU English proficiency test: Reliability related issues. *Journal of Humanities (16)*, 55-73.

Kline, T.J.B. (2005). *Psychological Testing*: *A Practical Approach to Design and Evaluation*. Sage Publications.

Kohli, N., Koran, J. & Henn. L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement, 75*(3), 389-405.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2015). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Loe, A. (2021). *Intro to IRT*. Available at https:// aidenloe.github.io.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Malec,W. & Krzemińska-Adamek, A. (2020). A practical comparison of selected methods of evaluating multiple-choice options through classical item analysis. *Practical Assessment, Research, and Evaluation: Vol.25*, Article 7. Retrieved from https://scholarworks.umass.edu/pare/vol25/iss1/7

Mehrens, W. A., & Lehmann, I.J. (1991*). Measurement and evaluation in education and psychology (4th ed)*. Belmont, CA: Wadsworth.Thomson Learning.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement, (3rd ed. pp*. 13-103). New York: American Council of Education and Macmillan Publishing Company.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R.W. Robins, R.C. Fraley, & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York: Guilford.

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment. The methodological, psychological and policy contributions of ETS* (pp. 19–46). Springer Open. http://dx.doi.org/10.1007/978-3-319- 58689-2_2.

Mousavi, A. (2009).*An encyclopedic dictionary of language testing*. Rahnama Press, Tehran.

Nguyen, T. H., Han, H. R., Kim, M.T. & Chan, K.S.(2014).An introduction to item response theory for patient-reported outcome measurement. *Patient, (7)*, 23-35. Springer. https://doi.org/10.1007/s40271-013-0041-0

Noori, M. & Hosseini Zadeh, S. (2017). The English Proficiency Test of the Iranian Ministry of Science, Research, and Technology: A Review. *International Journal of English Language & Translation Studies. 5*(3). 21-26.

Osterlind, S.J. (1983). *Test item bias*. Beverly Hills: Sage Publications.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed response, performance, and other formats (2nd ed.).* Boston, MA: Kluwer Academic.

Ockey, G.J. (2012). Item response theory. In G. Fulcher and F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp.336-345). Routledge.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Boston, MA: Allyn & Bacon.

Rizopoulos, D. (2018). *ltm: An R package for latent trait models under IRT*. Retrieved from https://github.com/drizopoulos/ltm.

Robitzsch, A. (2019). *sirt: Supplementary item response theory models. R package version 3.4-64.* Retrieved from https://CRAN.R-project.org/package=sirt

Sahrai, R. & Mamagani , H. (2013). The assessment of the reliability and validity of the MSRT proficiency test. *The Educational Assessment Journal, 10*(3), 1-19 [In Persian].

Salehi, M. (2011). On the construct validity of the reading section of the University of Tehran English Proficiency Test. *Journal of English Language Teaching and Learning, (222)*, 129-159.

Sawaki, Y. (2013). Classical test theory. In A. Kunnan (Ed.), *The companion to language assessment. Vol. 3*. Hoboken, NJ: Wiley Blackwell.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education, 19* (pp. 405-450). Washington, DC: American Educational Research Association.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement Issues and Practice (16),* 8–14.

Tsutakawa, R. k, & Johnson, J.C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, (55)*, 371-390.

Wiersma, W., & Jurs, S. (1990). *Educational measurement and testing*. Needham Heights, MA: Allyn and Bacon.

Wright, R.J. (2008). *Educational Assessment: Tests and measurements in the age of accountability*. Sage Publications.

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

Yu, C. H. (2010). *A simple guide to the item response theory (IRT) and Rasch modeling*. Retrieved from http://www.creative-wisdom.com.

Zimowski, M., Muraki, E., Mislevy, R. J., Bock, R. D. (2002). *BILOG-MG [Computer software]*. Lincolnwood, IL: Scientific Software International.