


REVIEW

Open Access



Test review of Iranian English language proficiency test: MSRT test

Ali Khodi^{1*} , Logendra Stanley Ponniah¹, Amir Hossein Farrokhi² and Fateme Sadeghi²

*Correspondence:
Alikhodi92@gmail.com

¹ Taylor's University, Subang Jaya,
Malaysia

² University of Neyshabur,
Neyshabur, Iran

Abstract

The current article evaluates a national English language proficiency test known as the “MSRT test” which is used to determine the eligibility of candidates for admission to and completion of higher education programs in Iran. Students in all majors take this standardized, high-stake criterion-referenced test to determine if they have obtained the minimum English proficiency level and can be graduated. The present paper seeks to examine the test and its psychometrics characteristics due to the significance of such a high-stakes examination that may have social and long-lasting effects on the participants. It is claimed that the test measures participants’ “knowledge of language” for communication rather than their “knowledge about language” in a constrained context. As a result, the test dimensionality and validity are up for debate. It was found that fundamental revisions in terms of test format and content are required to improve the test quality. The current study examined the areas that are yet unexplored and attempts to describe the MSRT assessment comprehensively.

Keywords: Evaluation, Iranian university English exam, Language test, MSRT examination, Test review

Introduction

The MSRT test, organized by the “Student Affairs Organization” affiliated with Iran’s Ministry of Science, Research and Technology, tends to evaluate the English language proficiency of the candidates. It is administered approximately ten times per year in major cities, with results announced through the exam portal within 3 to 4 days. Previously known as MCHE (Ministry of Culture and Higher Education), MSRT stands for Ministry of Science, Research and Technology.

The MSRT test, similar to the TOLIMO test, is considered a necessary tool in the evaluation process for doctoral applicants in Iran. It is sometimes referred to as internal TOEFL or doctoral TOEFL. While some institutions require passing grades on these exams, they are simpler and less expensive than TOEFL and IELTS, despite being referred to as a “little TOEFL test.” Farsi resources have been available for this exam for years, eliminating the need for applicants to use materials in their native language.

The MSRT test is taken by master’s and PhD students to assess their language competence, with many candidates being doctoral candidates due to the language

test requirement for the doctoral exam. The Ministry of Science, Research, and Technology, has been organizing the MSRT exam.

The exam booklet consists of three sections: 30 listening questions, followed by 30 grammar and 40 reading questions. Incorrect answers do not impact correct answers, but multiple responses to a question result in a zero score. The test is scored out of 100, with one point awarded for each correct answer. PhD students typically require a minimum score of 50 to pass, while a score of 40 is ideal for bachelor's degree candidates seeking admission to graduate studies.

Retaking the exam is allowed twice for better scores, except for candidates who score above 75, as retakes are restricted for a year. The exam's degree is authorized by the Ministry of Science and recognized by approved universities. However, the certificate is valid for two years and cannot be used for admission to international universities. The MSRT test assesses language proficiency through listening, reading, and grammar items. However, it may not provide a comprehensive measure of language proficiency. Language ability, as is defined in principle in this test, includes producing conversational content and maintaining coherence in the written format. The test's limitations in evaluating these aspects raise concerns about its reliability. Investigating the psychometric aspects will determine the test's capacity to assess language proficiency.

In terms of the scoring procedures, it should be mentioned that the MSRT lacks negative points for wrong answers, leading to chance achievement in scores due to guessing. However, deducting points, in special tests that apply it, may make test-takers cautious, resulting in unanswered questions. Although this test seems to be generally reliable, valid, and practical, with easy accessibility, computerized scoring, and reasonable fees. However, it does not assess speaking and writing skills, lacks real-world relevance, and has inappropriate administration conditions.

Despite the fact that the test is practical, convenient accessible, user-friendly with computerized scoring, and cost-effectiveness, it is imperative to approach the assessment of its validity and reliability with utmost caution (Tadayon and Khodi, 2017). It is essential not to underestimate the fundamental importance of ensuring that the test effectively measures the intended constructs and consistently yields dependable results. Thus, the results of several research studies conducted to evaluate these critical aspects in the last 5 years have been summarized in Table 1.

As indicated in the literature, the MSRT test exhibits several shortcomings that warrant attention. These include a lack of sufficient discriminating power, gender bias in its results, inappropriate listening conditions, the need for standard score determination, and imbalanced item difficulty. A detailed analysis of the test will be presented.

Test formats

The test is structured into three distinct sections, each targeting different aspects of English language proficiency. Listening comprehension, grammar (structure and written expression), and comprehension (reading and vocabulary) of the English language all together measure the degree of English competence.

Table 1 Summary of studies on MSRT Test (last 5 years)

Study title	Year	Summary of findings
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 (Ghahraki, Tavakoli, & Ketabi, 2022)	2022	This study states that the MSRT English proficiency test has important consequences but it is difficult for test-takers because it accommodates some items that do not function effectively or work negatively. Also, the findings suggest the need for accountability and quality assurance in the test. The findings in this study showed some items lack discriminating power
Examining differential item functioning (DIF) for Iranian EFL test takers with different fields of study (Rashvand Semiyari, & Ahangari, 2022).	2022	This study examines differential item functioning (DIF) in the MSRT test, focusing on two groups of test-takers from different academic majors. It was found that the results indicate that science students outperformed humanities students, particularly in the structure and written expressions (SWE) and reading comprehension (RC) sections. The study also recommends that other variables such as gender may bring about bias in the assessment results
Exploring the shortcomings of the Iranian MSRT English proficiency test (Ghorbani, Abbassi, & Razali, 2021)	2021	This study examined the experience of MSRT test-takers. It suggests that MSRT fails to assess productive speaking and writing skills, and this fact results in an underrepresentation of the construct it aims to measure (language proficiency). Moreover, it accentuates the inappropriate listening conditions in some testing centers which ends in bias
Comparison of ANGOF-based IRT method and bookmark method for the standard setting of MSRT language test (Jalalizadeh et al., 2019)	2019	This study compares the Angof-based IRT method and the Bookmark method. The results indicated that both methods yield higher cut-off scores than those determined by the Ministry of Science. The research suggests a need for revising the standard score determination of the MSRT test
A qualitative investigation of factors affecting a preparation course for MSRT: a grounded theory (Heshmatifar, Zareian, & Davoudi, 2018)	2019	This study used grounded theory and in-depth interviews to identify influential factors for enrolling in MSRT test preparation courses. A theoretical model was developed
The application of G-theory on MSRT (MCHE) score dependability; variability due to persons, gender, subject fields, sections, and items (Ghorbani, Abbassi, & Razali, 2021)	2019	This study examined the factors contributing to the dependability of MSRT test scores, including persons, gender, subject fields, sections, and items. The results showed high reliability and the importance of item difficulty and section scores, while gender had minimal impact. The findings can inform test development and research

Listening comprehension

This segment consists of 30 questions in total and features discussions between individuals with American accents. Following each chat and a brief pause, several questions related to that conversation are posed. You have 15 min to respond to these questions. In order to be better prepared to perform when answering the questions of the listening part, applicants can use numerous resources accessible in the market and online.

Grammar

This part consists of 30 questions, divided into two categories. In the first category (15 questions), you must provide the correct response to questions in which a part of a phrase is blank. You will come across questions in the second category, known

as error detection questions (15 questions), where the statement contains an error, and you must pick the erroneous response to represent the correct answer. You have 20 min to complete these questions. Most of the MSRT grammar questions are from the TOEFL book by Longman.

Reading comprehension

There are 40 questions in this section, which typically includes four academic texts for which you must provide answers. This process will take 40 min to complete. The subjects in the reading part of the TOEFL Longman book are frequently referenced in the comprehension section's questions. The reading exam segment often contains several vocabulary questions. The 504 and Essential Words for the TOEFL books are the major sources for vocabulary questions.

Test qualities and psychometric aspects

Validity and reliability

The question "Does a test measure what it is designed to assess?" is often asked when a test is administered. It refers to the concept of validity, which was mentioned by Lado (1961). For an examination's results to be taken into account, comprehended, and correctly applied, it must possess validity (Heaton, 1975, p. 153). Validity technically is defined as "the degree to which we may interpret a particular test score as an indicator of the abilities or constructs, we aim to measure" for the test (Bachman & Palmer, 1996, p. 21). As an academic examination intended to assess candidates' proficiency, the MSRT test should encompass certain characteristics, including validity and reliability. Validity for the current exam refers to assessing what the test is intended to measure, but we think that this outdated definition should also include the social validity (Chalhoub-Deville, 2016) aspect of the interpretations of the results (Messick, 1980). If it is claimed that the test examines the communicative ability of the learners and candidates, it should function accordingly. If it assesses something else while the students focus on the claimed aspect, they may lose their chance to pass the test, and ultimately, they will face problems in their education. Thus, the social aspect of validity is violated. We strongly believe that no test could be regarded as valid outside of the intended application and context (Messick, 1989). Thus, various aspects relating to the test's validity are discussed, clarified, and meticulously considered we go along (Messick, 2013).

As indicated by the statistical analyses and emphasized in the literature, the MSRT designers were not highly successful to accommodate the concept of language proficiency into the form of written questions. Missing other language sub-skills such as writing (Zabihi, Mehrani-rad, Khodi, 2019), and speaking justifies that the test questions are very limited to accommodate all dimensions of a complex concept such as language proficiency (Karami and Khodi, 2021). Speaking and writing were not taken into account in the MSRT test due to the difficulty in assessing the test-takers fairly, difficulty of operationalization of the performance in these skill, or their low level of importance for the examiners ; however, it is worth considering the possibility of exploring alternative formats or leveraging computer-based technologies to elevate the overall quality of the test. Accordingly, the subjectivity involved in rating these productive skills (Bloom, 1998), as the main reason for removing them, would be eliminated. The addition of these skills is

raised here due to the fact that in order to be able to say, with a relatively high percentage of certainty, that the participant is capable of using the language, a test like MSRT should be able to measure the quality of the test taker's ability to communicate and maintain a conversation or to present and develop his or her ideas through writing, too.

One notable issue with the MSRT test is its incomplete representation of construct validity, as there exists a gap between the so-called curriculum and the actual test (Messick, 1989). The sources necessary for studying prior to the test are similar to the sources of the TOEFL or IELTS tests. However, the nature of the questions differs fundamentally between these tests (Bachman, 1990). It could be claimed that no 100% determined source is available for the MSRT test except those grammar books. While the test's structure suggests multiple dimensions, the scoring treats all dimensions equally without considering item difficulty (Alavi et al., 2021). In order to increase the test discrimination weighted scores should be assigned to the items while we see all sub-skills are considered the same (Bachman, 2000; Khalilzadeh & Khodi, 2021). Furthermore, it has been found that the testlet model has the highest ability to explain the factor structure of language proficiency (Alavi et al., 2021); however, the scoring procedure of the MSRT test does not make use of any of validation models and techniques such as testlet or bifactor, higher-order models in item response theory (Khodi, Alavi, & Karami, 2021).

Overall, it seems that the concept of academic language proficiency the MSRT exam is going to operationalize somehow differs from the general concept of language proficiency defined in the literature because they try to define it in the form of the reading, grammar, and vocabulary skills that are essential for a PhD student to succeed at a university. It's necessary to say that this type of operationalization of language proficiency is valid locally because typically the format of the tests that are administered at the national level almost corresponds with the expectations that exist in the context. A careful examination of the items depicts that in order to be able to answer the questions not only knowledge about language was needed but also some levels of memorization were required too; the implication of such a fact is the questionability of the content validity for the evaluators of the test function. One should remember that generally the paper-based or electronic tests that seek to find out the language proficiency of test takers also struggle with the fact that they are not fully competent to assess the communicative competence of the participants.

The consistency of measurement is a concern that is connected to test usefulness, generalizability, and reliability (Khodi, Khezerlou, & Sahraei, 2022). Given the significance of the test results, the developers of the MSRT examination should strive to fulfill the need for test reliability. As the format of the test implies, the quantity of the items is a reflection of the consideration being given to the objectivity of measurement while some aspects such as uneven distribution of items in subsections, equal weight for items and not considering negative points for wrong answers are among the existing concerns of validity. When we delved into the analysis of the test's internal consistency, we discovered that there are unequal levels of reliability across the different sections. One reason for this disparity is the varying number of items in each section. It seems that insufficient control over the difficulty level of the test items contributes to this discrepancy. But that's not all; another factor that adds to the reliability issues of the MSRT test is

the ambiguity of the evaluation criteria. These factors combined make it challenging to establish a consistent and reliable assessment.

Optimization and generalizability analysis of the test

When a test is administered, a question arises: Do the test results accurately reflect the test takers' true ability, or are there external factors that influence the results, apart from the test takers' actual abilities? The answer to such a question could be found in classical test theories but in generalizability analysis of the test results. Usually in the administration of high-stakes and standardized tests the test administration context, the number of items, and the same scoring procedure have kept stable in order to eliminate any source of construct irrelevant variance (Khodi, 2021). This could be problematic due to two reasons: first, there is no intention to revise and improve the test quality by the test administrators and developers; second, keeping these conditions stable without examining the relative and absolute contribution they have would not be really insightful and helpful in reliability assessment. Particularly, for the MSRT test, as mentioned earlier, the optimization analysis revealed that an equal number for each subsection is needed and the increase in the total number of items would result in the improvement similar results were also found previously for the Konkur examination in Iran.

A generalizability analysis conducted in this study on a sample of 500 examinees revealed that 78% of the overall variation can be accounted for by individuals, indicating a high level of test reliability. We looked at the possible impact of age and major on performance. One interesting point about the g-analysis was the high index of the grammar section. For this part, the lowest possible number of items could be used and it showed no significant interaction with other variables. The total g-index was 0.78 and it improved to 0.90 once the total number of items increased by 10.

Test dimensionality and dependability of results

The MSRT test format and content highlight that basically the construct of measurement, that was language proficiency here, was considered to be multidimensional by the test developers which corresponds with the existing literature about the nature of language proficiency tests. Nevertheless, with regards to the other studies that have been conducted all language sub-skills do not contribute to the communicative competence of the learners similarly; thus, a relative and flexible scoring procedure is needed to be assigned.

Weighted scores to items based on the level of difficulty.

With regard to the MSRT's format and significance, participants may spend more time on grammatical structures, vocabulary, and reading than on writing, speaking, and listening exercises (Farhady et al., 1994; Ghorbani, 2012) the test; this is due to the underestimation of some language skills and has been accentuated in IRT analysis of the results and necessitates a fundamental change in the structure of the test. For instance, the weight dedicated to each subsection should be relative and the communicative ability should be the ultimate goal of items.

The effects of this exam touch the entire society, not only the PhD students and their families, who also place great importance on it. For instance, some language skills whose acquisition is easier will be more popular and the important aspect of language learning that is communicative ability will be under-valued.

Test bias and fairness

After carefully considering and conducting an Item Response Theory (IRT) analysis, it was discovered that certain items in the study exhibit unequal functioning for participants based on factors such as gender, major, and age. These differences and biases can be explained. For instance, grammar items were found to favor male participants, indicating a disparity in memorization abilities where female learners benefit less. When such biases occur, it indicates differential item functioning, which should be addressed by test developers as it significantly impacts test fairness. Furthermore, it was observed that the topics utilized in the listening and reading sections were not neutral, leading to unequal advantages for different groups. To ensure fairness, it is crucial to select topics that are both unbiased and inclusive, promoting equal performance among all participants. One possible solution is to conduct a pilot study during the test’s development phase, allowing developers to identify and eliminate items that exhibit differential item functioning.

However, it is unrealistic to completely eliminate a certain number of items since it would disrupt the test’s integrity. Instead, an alternate suggestion is to incorporate standard topics commonly used in tests like IELTS and TOEFL, which are known for their fairness and universality. Addressing differential item functioning and ensuring fairness in test administration can be achieved through careful item selection, topic neutrality, and the adoption of established standards in test development. These measures will help create an equitable testing environment for all participants (Fig. 1).

DIF class specification is: DIF=@GENDER

PERSON CLASS	Obs-Average	Exp-Measure	DIF-MEASURE	DIF-S.E.	PERSON CLASS	Obs-Average	Exp-Measure	DIF-MEASURE	DIF-S.E.	DIF-CONTRAST	JOINT S.E.	Rasch-Welch t	d.f.	Prob.	Mantel-Haenszel Chi-squ	Size Prob.	Active Slices	ITEM Name
1	.00	-.43	.38	2	.00	-.48	.36	.05	.53	.09	70	.9258	.0610	.8849	.01	13	1	ITEM1
1	-.03	-.43	.38	2	.03	-.75	.37	.32	.53	.60	70	.5518	.0081	.9282	.19	13	2	ITEM2
1	.07	-1.43	.45	2	-.07	-.61	.37	-.82	.58	-1.42	69	1.601	1.1864	.2761	-.95	13	3	ITEM3
1	.01	-.89	.40	2	-.01	-.75	.37	-.14	.55	-.25	70	.7995	.0580	.8096	-.04	13	4	ITEM4
1	-.01	-.01	.37	2	.01	-.10	.35	.09	.51	.18	70	.8570	.0175	.8947	.08	13	5	ITEM5
1	.03	1.44	.41	2	-.03	1.80	.43	-.35	.60	-.59	70	.5543	.0514	.8207	-.34	13	6	ITEM6
1	.06	.68	.37	2	-.06	1.30	.38	-.62	.54	-1.16	70	.2509	.8922	.3449	-.71	13	7	ITEM7
1	-.05	-.14	.37	2	.05	-.61	.37	.47	.52	.89	70	.3768	.2877	.5917	.43	13	8	ITEM8
1	.00	.01	.37	2	.00	.01	.35	.00	.51	.00	70	1.000	.0588	.8084	-.25	13	9	ITEM9
1	-.03	.27	.37	2	.02	.02	.35	.24	.51	.48	70	.6336	.1117	.7382	.34	13	10	ITEM10
1	-.04	1.81	.45	2	.04	1.30	.38	.51	.59	.86	70	.3902	.2320	.6301	-.41	13	11	ITEM11
1	-.01	.13	.37	2	.01	.02	.35	.11	.51	.21	70	.8339	.0528	.8183	-.02	13	12	ITEM12
1	-.04	-.58	.39	2	.04	-1.04	.39	.47	.55	.85	70	.4008	.1492	.6993	-.44	13	13	ITEM13
1	.08	-.89	.40	2	-.07	-.10	.35	-.79	.53	-1.48	70	.1439	1.2300	.2674	-.79	13	14	ITEM14
1	.03	-1.89	.51	2	-.03	-1.37	.42	-.51	.66	-.78	69	.4399	.0370	.8475	-.41	13	15	ITEM15
1	-.02	-1.89	.51	2	.02	-2.28	.55	.40	.75	.53	70	.5983	.0329	.8561	.41	13	16	ITEM16
1	.05	.27	.37	2	-.05	.76	.36	-.49	.51	-.96	70	.3423	.2781	.5979	-.43	13	17	ITEM17
1	-.03	2.26	.50	2	.03	1.80	.43	.46	.66	.70	69	.4864	.1016	.7499	.44	13	18	ITEM18
1	-.07	.13	.37	2	.06	-.48	.36	.61	.52	1.18	70	.2436	.5055	.4771	.54	13	19	ITEM19
1	-.04	2.02	.47	2	.04	1.45	.40	.57	.61	.92	69	.3591	.3835	.5357	.64	13	20	ITEM20
1	.07	-1.24	.43	2	-.06	-.48	.36	-.76	.56	-1.35	69	.1803	1.0248	.3114	-.75	13	21	ITEM21
1	-.12	.54	.37	2	.11	-.61	.37	1.15	.52	2.20	70	.0309	3.7710	.0521	1.33	13	22	ITEM22
1	.12	-.43	.38	2	-.11	.63	.35	-1.06	.52	-2.04	70	.0450	1.8192	.1774	-.86	13	23	ITEM23
1	.09	-.73	.39	2	-.08	.14	.35	-.87	.53	-1.66	70	.1011	.9414	.3319	-.67	13	24	ITEM24
1	-.12	1.44	.41	2	.11	.26	.35	1.17	.54	2.19	69	.0321	5.0864	.0241	2.02	13	25	ITEM25

Fig. 1 Differential item functioning

Test difficulty and discrimination

Upon examining the difficulty of items using IRT theory, it was discovered that certain items have an item difficulty index that deviates significantly from the average language proficiency level of the participants. Specifically, it was observed that the majority of grammar items were deemed easy for participants, irrespective of their gender, major, or age. In contrast, the majority of items in the listening section were found to be excessively challenging, resulting in a high number of unanswered questions. The ease of the grammar items can be attributed to the fact that all participants have studied grammar extensively. However, the difficulty of the Listening section arises from the lack of preparation available to participants, particularly in terms of listening skills. The listening tracks used in the test are sourced from standard international exams, which are known to be challenging for local participants who have not undergone specific preparation courses for such exams. It is important to note that the difficulty level of the local test, the MSRT, cannot be directly compared to other internationally recognized exams due to their distinct objectives. The primary goal of the local test is to evaluate whether prospective undergraduate and postgraduate students can effectively utilize current English-language resources and perform well in academic settings. On the other hand, international tests not only assess learners' readiness for functioning in academic settings but also evaluate their compatibility with the society they will be a part of if they are English speakers. Consequently, international tests are primarily designed to assess overall language proficiency rather than focusing on specific language aspects. Analysis reveals that 17% of the items were considered very easy, while 20% were deemed very difficult. This indicates that only 63% of the test items effectively measure the learners' abilities, which falls short of an optimal level for reliable results analysis. The examination of item difficulty using IRT theory highlighted discrepancies in the test's composition. While the grammar section proved to be relatively easy for participants, the Listening section posed significant challenges due to limited preparation resources. Addressing these issues is crucial to ensure the accuracy and validity of the test results for a comprehensive analysis.

Usefulness and practicality, authenticity, and interactivity

A test is created and administered for its usefulness, which is its most important attribute (Bachman & Palmer, 1996). The link between the resources that will be needed for the test's design, development, and usage, and the resources that will be available for these activities, is referred to as practicality (Bachman & Palmer, 1996, p. 39). Speaking and writing, which are the two most important abilities, are not examined in the MSRT test; therefore, the test is already unable to gauge the participants' productive abilities. Additionally, there is significant doubt regarding the test's outcomes. There will be practical restrictions taking into account time allotment for administering and scoring, the qualifications and availabilities of test administrators and scorers, and financial considerations even if the speaking and writing sections, which are the most problematic in constructing validity, are tested using a computer.

The short scoring time in this case is the biggest issue that both choices share. The scoring criteria for the speaking and writing portions will need to be discussed in great detail in order for the scorers to evaluate the exam fairly and impartially. However, the

MSRT score report must be released within 3 to 4 days following the exam in order for university admissions to proceed as planned. As a result, the MSRT exam is not very useful or practical and does not always address the issues with the test qualities mentioned. But it is not impossible either. By creating short-answer questions in the speech and writing portions utilizing recording equipment or computers, the MSRT test can address the poor practicality issue. The validity of the test will also be increased by having speaking and writing portions.

The degree of correlation between a test and real language usage tasks is known as authenticity (Bachman & Palmer, 1996, pp. 23–25). According to Morrow (2012), the test items should contain real-life texts from the perspective of authenticity. As was already stated, speaking and writing abilities are never evaluated in the examination. Because of this, the MSRT examination's authenticity is poor and it cannot deal with the issue.

The test items should also trigger the unique test-takers' qualities, such as language proficiency, metacognitive methods, and topical knowledge, for higher interactivity (Bachman & Palmer, 1996, p. 25). The MSRT exam's interactivity is in doubt because basic memorization may be particularly pertinent. The MSRT exam has low interactivity since it only partially uses the learner's language skills and because little metacognitive thinking and topical knowledge are demanded.

Conclusion

MSRT acknowledges the strategic arrangement of provided materials and comprehensive coverage of resource books. Despite its deconstructive aspect, the impact on teaching and learning can be restored through the integration of new source materials and revised priorities. The challenges related to the exclusion of writing and speaking skills can be addressed by introducing additional information into teaching resources and redesigning the exam. A new test format can also address the issue of poor practicality. Enhancing reliability can be achieved by incorporating clear assessment criteria. To ensure authenticity and interactivity, it is important to include commonly used grammar in everyday discourse. The speaking and writing sections of the exam should be designed using computer-based tools and/or recordings to uphold test validity. Although there may be some irrelevant variance, incorporating weighted scores and considering participants' educational backgrounds can help reduce it. Once there is consensus on the construct being tested and the utilization of new technology in the assessment process, the remaining issues can be addressed.

The mismatch between the nature of test items and the item format is a key reason for some participants' failure. The comprehensive evaluation of MSRT reveals that despite contextual and content alignment, the test's nature does not effectively predict participants' actual language proficiency as it lacks the inclusion of crucial language abilities. Encouragingly, there is a sufficient body of research reviewing MSRT and providing suggestions for improvement, considering its social and long-term impacts on Iranian society. The MSRT exam has drawbacks that compromise its effectiveness. It lacks construct validity by excluding speaking and writing skills, raising concerns about accurately assessing communication abilities. Content validity is also an issue as some items prioritize memorization over communication competence. Indirect assessment methods hinder reflecting true competence. Reliability problems arise from uneven question

quantity, equal weight for incorrect options, and skill interference. These limitations undermine validity, reliability, and generalizability.

To address these drawbacks, include speaking and writing sections to assess communication abilities accurately. Improve content validity by covering comprehensive language skills. Use direct assessment methods for accurate measurement. Enhance test design by balancing questions and weighting item difficulty. Provide comprehensive training for raters. Involve stakeholders and consider social validity to enhance practicality and relevance. Lastly, institutions should establish more inclusive admission practices that consider students' performance on the MSRT foreign language examination. Overall, this reconsideration will benefit the educational system and enhance long-term learning motivation.

To mention some of the challenges and propose potential solutions it can be stated that one significant issue is the unequal importance assigned to different items and subsections within the exams. This disparity may lead to an undervaluation of essential aspects such as communicative ability while placing excessive emphasis on areas like grammar, which candidates find relatively easier. To address this, we recommend implementing a scoring procedure that dynamically reflects the goals of the exam. This approach would grant greater significance and more diverse assessment methods for measuring communication skills, aligning the evaluation process with the exam's intended objectives.

Another concern lies in the lack of a clear, transparent explanation of the exam's goals. Many students participate in the exam solely with the aim of meeting the minimum cutoff score, without fully comprehending the broader language proficiency development that the exam should foster. To tackle this issue, we propose integrating additional productive skills, like writing or speaking, into the test. By doing so, we can transform the exam into a comprehensive tool that promotes genuine learning rather than rote memorization, ultimately yielding more positive outcomes. Furthermore, the problem of question repetition needs to be addressed. Currently, a significant percentage of items are repeated from previous exams, compromising the authenticity and fairness of the assessment process. To rectify this, it is crucial to introduce a wider range of questions and ensure rigorous validation and development of test materials. Drawing inspiration from established systems such as IELTS or TOEFL, we can enhance the quality and diversity of questions while adhering to robust validation practices. The timing and administration of the exam pose additional concerns. In some language centers, the inadequate presentation of sound during the listening section can undermine candidates' ability to demonstrate their language skills effectively. Moreover, the excessive number of items in the exam necessitates a reduction. Conducting a comprehensive item analysis and power analysis will help determine the optimal number of items for each subsection. Additionally, the correlation between different subsections reveals issues with the exam's structure and dimensionality, calling for a thorough reassessment and restructuring.

In conclusion, addressing the concerns related to the exam's fairness, validity, construct value, content, administration, and structure is vital to ensure a robust and accurate assessment of language proficiency. By implementing the proposed solutions, we can create an exam that provides reliable and comprehensive insights into candidates' abilities and growth.

Abbreviations

MSRT	Mistry of Science, Research and Technology
MCHE	Ministry of Culture and Higher Education
TOLIMO	Test of Language by the Iranian Measurement Organization

Acknowledgements

To all those editors who read and revised the papers.

Authors' contributions

The authors made the same contribution. All authors read and approved the final manuscript.

Funding

We received no funding.

Availability of data and materials

The data will be available to third parties upon request and justification.

Declarations**Competing interests**

The authors declare that they have no competing interests.

Received: 18 October 2023 Accepted: 5 December 2023

Published online: 08 February 2024

References

- Alavi, S. M., Karami, H., & Khodi, A. (2021). Examination of factorial structure of Iranian English language proficiency test: An IRT analysis of Konkur examination. In *Current Psychology*, 42(10), 8097–8111. <https://doi.org/10.1007/s12144-021-01922-1>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, 17(1), 1–42.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). London: Oxford University Press
- Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences*, 2(2), 67–73.
- Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, 33(4), 453–472.
- Farhady, H., Jafarpur, A., & Birjandi, P. (1994). *Testing language skills: From theory to practice*. SAMT Publications.
- Ghahraiki, 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. *Two Quarterly Journal of English Language Teaching and Learning University of Tabriz*, 14(29), 1–26.
- Ghorbani, M. R. (2012). Controversy over abolishing Iranian university entrance examination", *Asian Education and Development Studies*. 1(2), 139–152. <https://doi.org/10.1108/20463161211240115>
- Ghorbani, M. R., Abbassi, H., & Razali, A. B. M. (2021). Exploring the Shortcomings of the Iranian MSRT English Proficiency Test. *Pertanika Journal of Social Sciences & Humanities*, 29(S3), 115–132
- Heaton, J. (1975). *Writing language tests*. Longman.
- Heshmatifar, Z., Zareian, G. R., & Davoudi, M. (2018). A qualitative investigation of factors affecting a preparation course for MSRT: a grounded-theory. *Journal of Foreign Language Research*, 8(1), 27–56.
- Jalalzadeh, M., Delavar, A., Farokhi, N., & Askari, M. (2019). Comparison of ANCOF-based IRT method and Bookmark method for standard Setting of MSRT language test. *Research in Teaching*, 7(4), 69–49.
- 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.
- Khalilzadeh, S., & Khodi, A. (2021). Teachers' personality traits and students' motivation: a structural equation modeling analysis. *Current Psychology*, 40(4), 1635–1650.
- Khodi, A. (2021). The affectability of writing assessment scores: a G-theory analysis of rater, task, and scoring method contribution. *Language Testing in Asia*, 11(1), 30.
- Khodi, A., Alavi, S.M. & Karami, H. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Lang Test Asia*, 11, 14. <https://doi.org/10.1186/s40468->
- Khodi, A., Khezerlou, H., & Sahraei, H. (2022). Dependability and utility of using e-portfolios in assessing EFL learners' speaking proficiency. *Computer Assisted Language Learning*, 1–23
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. Bristol: Longmans, Green and Company
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11
- Messick, S. J. (2013). Alternative modes of assessment, uniform standards of validity. In *Beyond multiple choice*. Psychology Press. pp. 59–74.
- Morrow, K. (2012). Communicative language testing. The Cambridge guide to second language assessment. 140
- Rashvand Semiyari, S., & Ahangari, S. (2022). Examining differential item functioning (DIF) for Iranian EFL test takers with different fields of study. *Research in English Language Pedagogy*, 10(1), 169–190.
- Tadayon, F., & Khodi, A. (2017). Empowerment of refugees by language: Can ESL learners affect the target culture? *TESL Canada Journal*, 33(10), 129–137
- Zabihi, R., Mehrani-Rad, M., & Khodi, A. (2019). Assessment of authorial voice strength in L2 argumentative written task performances: contributions of voice components to text quality. *Journal of Writing Research*, 11(2), 331–355.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.