Journal of English Language Teaching and Learning

University of Tabriz

Volume 17, Issue 35, 2025

Investigating the Reliability of the Reading Module of the Iranian Ministry of Science, Research, and Technology's English Proficiency Test

Ali-Akbar Ariamanesh (Corresponding Author)

Assistant Professor, English Department, Science and Arts University of Yazd, Yazd, Iran. E-mail: aa.ariamanesh@fgn.ui.ac.ir

Mohammad Hajimohammadi

Assistant Professor, English Department, Science and Arts University of Yazd, Yazd, Iran. E-mail: hajimohammadi@sau.ac.ir

Fereshteh Dehghan-Manshadi

MA. in ELT, Science and Arts University of Yazd, Yazd, Iran. E-mail: dehghanmanshadifereshte@gmail.com

ARTICLE INFO:

Received date:
2025.01.05
Accepted date:
2025.03.19

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

Keywords:

MSRT Reading, Test Reliability, Topic Effect, Item Type, Test Bias

Abstract

As a national high-stakes test of English proficiency, MSRT needs further scrutiny of reliability. Thus, the present study aimed to investigate different sources of variation that may impact the MSRT test-takers' reading comprehension. Accordingly, some factors including the reading topics, item types, and participants' general proficiency were delved into based on the scores obtained from 60 MSRT prospective candidates. Upon administering a sample of the reading subtest taken from a recent version of MSRT, the collected data was dichotomously scored and then analyzed in terms of internal consistency, inter-correlations, and causal patterns. The yielded results showed an overall reliability of 0.86 for the reading module, while a moderate interrelationship was obtained amongst the passages (r = 0.47) as well as the item types (r = 0.44). Furthermore, the Mixed ANOVA results demonstrated that topic and item type significantly affected the reading performance, whereas the proficiency factor did not play a conspicuous role in distinguishing the participants' reading accomplishment. Both theoretically and operationally, the results reported through this study may stress the need for reconsidering the influence of such test-method facets as topic and item type in the MSRT reading subtest to upgrade the test's unidimensionality and fairness.

Citation: Ariamanesh, A. A.; Hajimohammadi, M. & Dehghan-Manshadi, F. (2025). Investigating the Reliability of the Reading Module of the Iranian Ministry of Science, Research, and Technology's English Proficiency Test. *Journal of English Language Teaching and Learning*, 17 (35), 01-14. DOI: 10.22034/elt.2025.65321.2737

Introduction

English language proficiency tests are crucial in the Iranian educational context, particularly for individuals seeking admission to postgraduate studies or applying for scholarships abroad. One of the important English language proficiency tests in Iran is the MSRT exam. The acronym stands for the Ministry of Science, Research and Technology, which was formerly known as the MCHE (Ministry of Culture & Higher Education) exam. This standardized test of English language proficiency is administered by the Student Affairs Organization affiliated with Iran's Ministry of Science, Research and Technology. The test is designed to assess the general proficiency of, especially, Ph.D. candidates in various majors.

MSRT is a paper-based exam and consists of 100 multiple-choice items. All the items have equal weight and the test is scored out of 100. In this test, guessing is not penalized and the total test duration is 110 minutes. MSRT comprises three subtests: Listening comprehension (30 items), structure and written expressions (30 items), and reading comprehension (40 items). As a widely used Iranian test of general English proficiency, the MSRT exam needs to meet the psychometric properties of validity and reliability (Bachman & Palmer, 2010). The present study was concerned with investigating the reliability of the reading comprehension section of this test. Reading comprehension assessment involves a number of strategies and techniques that are intended to show the extent to which a learner can read, understand, interpret, and analyze a text (Brown & Abeywickrama, 2018). In fact, reading comprehension is a complicated mental activity affected by a lot of factors (Hill & Liu, 2012). Amongst the various factors that affect second language reading test performance are text topic and item type.

Technical topics in reading comprehension assessment can be highly problematic since they are in favor of some groups of participants. Therefore, some participants score higher on the reading section and this higher score is not due to their reading ability but it is because of their previous knowledge on the topic. In second language assessment, any factor that influences test scores, except the construct under measurement, is regarded as a possible source of measurement error. Such measurement errors can limit the reliability and generalizability of test scores (Chalhoub-Deville & Turner, 2000). Item type, which is subsumed under test-method facets (Bachman, 1990), is another factor that affects test performance. In this regard, the MSRT reading section includes a variety of item types with different functions. The examinees are required to answer main idea, inference, content-based, vocabulary, information-accuracy, etc. questions. Some of the item types, explicitly those designed to elicit inference and main idea, may be subjective, which can adversely affect the respondents' performance. Another concern about the item types pertains to the extent to which test items measure the same underlying construct. If test items are not homogeneous in measuring a single underlying trait, the internal consistency of the test will be jeopardized (Fulcher, 2010).

In sum, the focus of the present study was on investigating how much the MSRT reading subtest is free of error, unbiased in terms of topic, and unidimensional in terms of measuring a single underlying construct.

1. Literature Review

Although the MSRT exam is considered to be a quite famous English general proficiency assessment tool in Iran, its features of what makes a good test have not received enough attention from relevant scholars. Amongst the quality criteria of a standard second language assessment instrument, reliability seems to be more fundamental because it pertains to the degree to which a test is free of errors, and thereby, able to yield consistent results (Bachman, 1990). Concerning the MSRT test, a number of studies have so far addressed its different technical aspects, including validity and reliability (Fallahian & Tabatabaei, 2015; Ghahraki et al., 2022; Khodi et al., 2024; Noori & Hosseini Zadeh, 2017; Rashvand Semiyari & Ahangari, 2022; Sahrai & Mamagani, 2013). As an example in case, Sahrai and Mamagani (2013) embarked on studying the validity and reliability of the MSRT exam. They found that the test had an acceptable level of reliability ($r > 0.7$), while the correlation coefficient between the grammar and reading comprehension sections was higher than the interrelationship between the listening and grammar as well as the listening and reading comprehension sections. These variations indicate the different components of the exam might not be very consistent.

An investigation by Fallahian and Tabatabaei (2015) concentrated explicitly on the construct validity of the reading comprehension section of the MSRT test through both qualitative and quantitative approaches. The observed results, however, could not clearly support the construct validity of the MSRT reading module because the triangulated data failed to yield unanimous results. In a similar but more comprehensive attempt, Noori and Hosseini Zadeh (2017) reviewed the MSRT test and its subsections. Having scrutinized the merits and demerits of the test, they contended that their study added evidence in support of an acceptable level of general reliability and validity of the MSRT exam. Using an item response theory (IRT) approach, Rashvand Semiyari and Ahangari (2022) examined the MSRT test through the differential item functioning (DIF) analysis. Accordingly, the reliability of the whole test turned out to be 0.85, whereas the reliability for the reading section was 0.73. Moreover, the results showed that those test takers whose academic major was in sciences had a better performance than the humanities students, especially, in structure and written expressions as well as the reading comprehension section. Certainly, the recent conclusion may raise concerns about the topic neutrality and fairness of the test under scrutiny.

In a more thorough study, Ghahraki et al. (2022) investigated the item properties of the MSRT exam using a 2-parameter IRT model. Item difficulty and item discrimination for the three sections of the test were assessed. The results of this study revealed that MSRT is a challenging endeavor for the test takers and a number of test items are either nonfunctioning or working negatively. However, all MSRT subsets met the unidimensionality assumption. Regarding the reading subsection, the analysis of difficulty and discrimination indices indicated that 20% of the items were either easy or very easy, 40% were moderate, and 40% were either difficult or very difficult. Moreover, 17.5% of reading items were nonfunctioning, 30% discriminated moderately, and 50% discriminated either highly or perfectly.

Khodi *et al.* (2024), in turn, evaluated the MSRT test, delving into its internal consistency. They discovered that there were unequal levels of reliability across the different components of the test. They also found that the reliability problems were related to uneven question

quantity, skill interference, equal weight for incorrect options, and the ambiguity of the evaluation criteria. Arguably, such limitations can undermine the validity, reliability, and generalizability of the test. Khodi et al. (2024) further perceived that the topics in the reading and listening sections were not neutral, which led to unequal advantages for different groups of test takers. This finding might be interpreted against test fairness in the MSRT exam.

Reviewing the past literature on the reliability of some other national and international English proficiency tests, explicitly their reading module, suggests several studies that can shed more light on the factors contributing to test reliability. In such a research work, Kiani and Haghighi (2006) investigated the reliability, test difficulty, and speededness of the English proficiency test of Tarbiat Modarres University of Iran. They concluded that although the exam was to some extent reliable, its reliability coefficients did not meet the levels that would be required for a high-stakes test. In a similar attempt, Karami (2012) examined the effect of factors such as test items and academic background on the dependability of the scores of the University of Tehran English Proficiency Test (UTEPT). Both classical reliability analyses and generalizability studies (G-studies) were conducted. The reliability estimates for the UTEPT were 0.90, 0.80, 0.77, and 0.78 for the whole test, grammar, vocabulary, and reading subtests, respectively. Further, the relative and absolute G coefficients were 0.86 and 0.84, successively. Moreover, the results of this study showed no bias due to the test-takers' background knowledge. Sticking to the principles of Generalizability Theory, Ahmadi Shirazi et al. (2019) investigated the effect of text and item types in the reading comprehension section of the Iranian PhD entrance exam. They obtained an internal consistency of 0.63 for the reading items, while their relative and absolute generalizability coefficients were in turn 0.65 and 0.64. Besides, the results indicated that item type played a significant role in the measurement precision of the test-takers' ability in the reading section. In contrast, text type did not significantly affect the test-takers' reading comprehension.

The two leading international proficiency tests of English, i.e. TOEFL iBT and IELTS, have also been studied in terms of validity and reliability. For instance, Hill and Liu (2012) investigated the interaction effect of background knowledge and language proficiency in the TOEFL iBT reading subtest. The test takers were classified into high- and low-proficiency groups based on their iBT total scores. It was found that the TOEFL reading texts were neither advantageous nor disadvantageous to those test takers who had prior knowledge of the topics presented by the passages, which held true for both proficiency groups. Likewise, Dewi et al. (2023) analyzed 20 TOEFL iBT reading comprehension questions in terms of their reliability, item and person fit, and difficulty level. Based on the results yielded by the one-parameter Rasch model, four items needed to be discarded since they were too easy or too difficult, but the other items were of good quality and met the standard requirements for a reliable test. As for the other high-stakes test, Hashemi and Daneshfar (2018) critically reviewed the IELTS test by focusing on its reliability, validity, and washback. They ultimately reported that the prestigious test carried an internal consistency of 0.88 for the listening and reading subtests. Focusing on technical vocabulary, Ashrafzadeh et al. (2015) examined the impact of background knowledge on L2 reading comprehension of Iranian medical students. The participants' comprehension was examined by an IELTS reading text that included sub-technical medical terms and a passage containing highly-technical medical terms. The study

ultimately concluded that background knowledge plays a distinctive role in second-language reading comprehension.

The above-reviewed literature, specifically on the MSRT reading subtest, provides valuable information on different factors affecting the test-takers' performance. Yet, it seems there should be a more precise investigation into how different text topics as well as item types influence the examinees' reading comprehension. Additionally, the degree of go-togetherness amongst the four reading passages and various item types may clarify the test's unidimensionality. Correspondingly, the current study attempted to address these issues with the aim of further delving into the reliability of the MSRT reading module.

Congruent with the specific aims of this investigation, the following research questions (RQs) were addressed to shed more light on the MSRT reading module's reliability.

RQ1: How much correlation exists amongst the four topics as well as the different item types in the MSRT reading module?

RQ2: Do the different topics and item types in the MSRT reading module significantly affect the test-takers' reading comprehension?

RQ3: Does overall proficiency have any significant effect on the test-takers' performance in the MSRT reading module?

2. Methodology

2.1. Design of the Study

This quantitative study adopted a causal-comparative (ex post facto) research design since none of the independent variables, i.e. proficiency, topic, and item type, was manipulated. In fact, the factors had already happened and their probable effects on the dependent variable, i.e. the test-takers' reading scores, were to be measured. According to Ary et al. (2019), ex post facto research is a method of teasing out the possible antecedents of events that have already happened and cannot be controlled or manipulated by the investigator. By the same token, Best and Khan (2016) contend that a causal-comparative study investigates the relationship amongst variables, where the independent variable has already occurred and the researcher has no control over it. Ultimately, the explored causal relationship is exploited to draw conclusions about the results.

2.2. Participants and Sampling

The participants employed in this study were 60 Iranian post-graduate students who were attending a preparation course, held by Yazd University Language Center, for Iranian national English proficiency exams (e.g., MSRT & TOLIMO). They were all adult Persian EFL learners of both genders from different university majors other than English. The participants were informed about the present research aims and almost all of them agreed to help with the data collection process. Accordingly, the sample was selected through a convenience sampling technique (Mackey & Gass, 2022). Table 1 summarizes the demographic information of the participants.

Table 1. *Demographic Background of the Participants*

No. of Participants	60
Gender	26 Males / 34 Females
Native Language	Persian
Academic Year	2023-2024

2.3. Instrumentation

The first instrument this study drew upon was the test of English proficiency administered by Yazd University Language Center. Actually, it was the final exam for the preparation course described in the previous section. The proficiency test, which had a score range of 0-100, comprised three subsections: Listening-speaking, grammar, and reading-writing.

The second and main instrument used in this study was a recent version of the MSRT English proficiency test. Since the focus of this study was on the reliability of the reading module of the MSRT exam, the test items related to the reading comprehension section were used to assess the participants' performance. The reading section contained four reading texts, each followed by 10 multiple-choice items. It should be noted that in this study two forms of the same test (Booklet A: Texts 1 to 4, & Booklet B: Texts 4 to1) were used to offset the effect of passage sequence.

2.4. Procedure

Upon the completion of the preparation course and based on the mean of their total scores (69 out of 100) in the final exam, the participants were divided into two proficiency levels. Correspondingly, those who scored from 69 downwards were classified as the lower proficiency group, while those who received 70 or more were classified as the higher proficiency group.

During two data-collection sessions, the reading comprehension section of the MSRT exam, which included four reading texts with an allocated time of 50 minutes, was administered to the participants who were real candidates for the national English proficiency tests in Iran. It should be clarified that the respondents were divided into two groups, where each group received a version of the test (Booklet A or B) randomly. More precisely, 48.3% of the examinees received Booklet A in which the sequence of the reading texts was the same as the original MSRT test, whereas 51.7% of the examinees received Booklet B with the reversed sequence of the reading texts. It is worth noting that in the process of data analysis by the SPSS software, the sequence of texts in Booklet B was primitively rebalanced so as to produce monolithic coded data.

2.5. Data Analysis

To address RQ1, Pearson Correlation was applied to expose the direction and magnitude of the interrelationships amongst the four topics and the various item types in the MSRT reading subtest. Regarding RQs 2 and 3, the Mixed Between Within Subjects Analysis of Variance (Pallant, 2020) statistical technique was conducted using the IBM SPSS (version 27) software. Deploying this analytical approach, it was possible to measure the influence of each within-subjects independent variable, i.e. passage topic and item type, as well as the effect of the between-subjects independent variable (proficiency) on the test-takers' reading performance.

In addition to these main effects, the results yielded by the Mixed ANOVA analyses revealed the interaction effects between the independent variables too.

3. Results

3.1. Normality of the Collected Data

As the normality of data is a prerequisite for many parametric statistical tests (Tabachnick & Fidell, 2013), the collected data was primarily tested via the skewness and kurtosis measures. Indeed, the four reading topics as well as the four item types went under the intended analysis to ascertain their normality (see Table 2).

Table 2. *Descriptive Statistics of the Four Topics and Four Item Types in the MSRT Reading Module*

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Geology	60	.5117	.25584	.111	.309	-.623	.608
History	60	.4900	.27414	.103	.309	-.999	.608
Anthropology	60	.5217	.20260	.061	.309	-.945	.608
Industry	60	.5983	.25278	-.059	.309	-1.057	.608
Inference	60	.4238	.21684	-.032	.309	-.606	.608
Paraphrase	60	.5636	.20210	.160	.309	-.412	.608
Info_Accuracy	60	.3417	.38501	.620	.309	-1.035	.608
Vocabulary	60	.5683	.24940	.012	.309	-1.152	.608

As Table 2 demonstrates, almost all of the skewness and kurtosis values were between -1 and +1. According to Hair et al. (2022), the skewness and kurtosis scores within the outlined range show fully acceptable normality of data.

3.2. Internal Consistency

Internal consistency concerns how much the items of a test are homogeneous in terms of measuring the same underlying construct. Referring to outstanding statisticians (Field, 2024; Pallant, 2020; Tabachnick & Fidell, 2013, amongst others), an instrument's internal consistency above 0.70 is considered acceptable, while values above 0.80 and 0.90 are more preferrable. In the present investigation, the overall internal consistency for 40 dichotomously-scored items of the MSRT reading subtest was computed via Cronbach's Alpha, which turned out to be 0.86.

Table 3. *Internal Consistency of the MSRT Reading Items*

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.865	.866	40

The obtained internal consistency, as shown in Table 3, demonstrates a magnitude of 0.866 based on standardized items, which is quite a high level of internal consistency for the reading section of the MSRT exam.

3.3. Correlation amongst the Four Reading Topics

To determine the degree of mutual relationship amongst the reading topics, the Pearson Product Moment correlation was applied and the pairwise coefficients amongst the topics were

computed by the SPSS software (Table 4). According to Pallant (2020), correlation coefficients between 0.10 and 0.29 are considered to be *small*, values between 0.30 and 0.49 are interpreted as *medium*, and those values from 0.50 upwards indicate a *large* correlation.

Table 4. *Correlations amongst the Four Topics of the MSRT Reading Module*

		Geology_ Mean	History_ Mean	Anthropology_ Mean	Industry_ Mean
Geology_Mean	Pearson Correlation	1	.558**	.521**	.396**
	Sig. (2-tailed)		.000	.000	.002
History_Mean	Pearson Correlation	.558**	1	.596**	.288*
	Sig. (2-tailed)	.000		.000	.025
Anthropology_Mean	Pearson Correlation	.521**	.596**	1	.497**
	Sig. (2-tailed)	.000	.000		.000
Industry_Mean	Pearson Correlation	.396**	.288*	.497**	1
	Sig. (2-tailed)	.002	.025	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

According to Table 4, the correlation coefficient between history and industry was 0.28, which shows a small correlation. The coefficients between geology and industry as well as between anthropology and industry were 0.39 and 0.49, respectively, which are interpreted as moderate. Further, the values between geology and history, geology and anthropology, and history and anthropology were 0.55, 0.52, and 0.59, respectively, showing a large interrelationship.

3.4. Correlation amongst the Four Item Types

Following the same statistical approach deployed for the topics, the bivariate correlation coefficients for the four item types in the MSRT reading subtest were also computed by the SPSS software to further analyze the reliability of the test.

Table 5. *Correlations amongst the Four Item Types of the MSRT Reading Module*

		Inference_ Mean	Paraphrase Mean	Info Accuracy _ Mean	Vocabulary _ Mean
Inference_Mean	Pearson Correlation	1	.354**	.484**	.501**
	Sig. (2-tailed)		.006	.000	.000
Paraphrase_Mean	Pearson Correlation	.354**	1	.424**	.433**
	Sig. (2-tailed)	.006		.001	.001
Info_Accuracy_Mean	Pearson Correlation	.484**	.424**	1	.485**
	Sig. (2-tailed)	.000	.001		.000
Vocabulary_Mean	Pearson Correlation	.501**	.433**	.485**	1
	Sig. (2-tailed)	.000	.001	.000	

** Correlation is significant at the 0.01 level (2-tailed).

As Table 5 displays, the correlation magnitude between inference and paraphrase, inference and information accuracy, paraphrase and information accuracy, paraphrase and vocabulary, as well as between information accuracy and vocabulary were successively 0.35, 0.48, 0.42,

0.43, and 0.48, which indicate a moderate correlation level. Moreover, the correlation between inference and vocabulary item types was 0.50, being interpreted as a large relationship.

3.5. The Effect of Topic and Proficiency

To explore the impact of topic and proficiency on the MSRT reading achievement, a mixed between-within-subjects analysis of variance (Mixed ANOVA) was run by the SPSS software. Reviewing the Box's Test of Equality of Covariance Matrices, no violation was found (Sig.= 0.338), which proved that the subgroups, i.e. the higher and lower proficiency participants, were homogeneous. Likewise, Levene's Test of Equality of Error Variances showed no violation as all yielded Sig. values were above the conventional Alpha level (0.05). This recent finding demonstrated the homogeneity of the four subtopics under analysis. The descriptive statistics for the MSRT reading topics produced through Mixed ANOVA are presented in Table 6.

Table 6. *Descriptive Statistics of the Four Topics of the MSRT Reading Module*

Topic	Proficiency	Mean	Std. Deviation	N
Geology	Lower	.4565	.24088	23
	Higher	.5459	.26204	37
History	Lower	.3783	.30742	23
	Higher	.5595	.22908	37
Anthropology	Lower	.5130	.21596	23
	Higher	.5270	.19671	37
Industry	Lower	.5826	.22290	23
	Higher	.6081	.27221	37

As Table 6 shows, the highest mean score (M = 0.60) belonged to the performance of the higher proficiency group in industry, whereas the lowest mean score (M = 0.37) was recorded for the lower proficiency group in history. The mean performances of the two groups of test takers in four topics of the MSRT reading section are depicted in Figure 1. It can be estimated that in anthropology and industry, proficiency did not cause a big difference between the proficiency groups, yet in history and geology, the between-groups factor exerted a more distinct difference between the lower and higher proficient test takers.

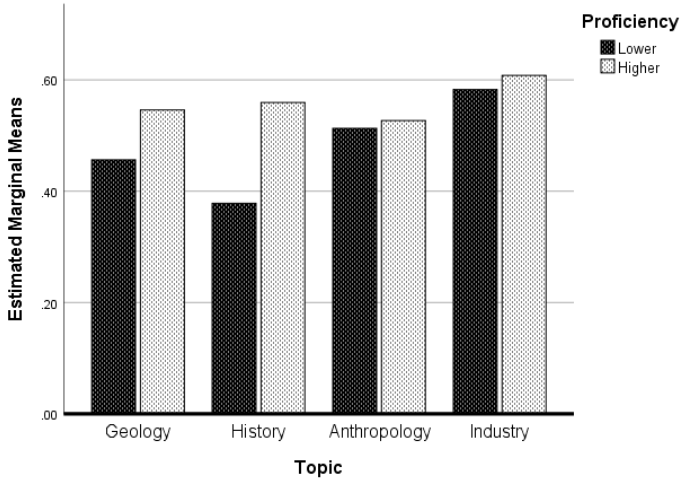

Figure 1. *Mean Scores of the Two Groups for the Four Topics of the MSRT Reading Module*

The Multivariate Tests table, in turn, revealed that the interaction effect of topic and proficiency (based on Wilks' Lambda) was significant at 0.042 level. The corresponding Partial Eta Squared value was 0.135, revealing a moderate effect size at the significant point of difference. The effect of the topic was also significant (Sig. = 0.029), carrying a large effect size (partial eta squared = 0.148). In contrast, the Tests of Between-Subjects Effects table indicated that proficiency did not play a significant role in distinguishing the two groups (P-value = 0.128).

3.6. The Effect of Item Type and Proficiency

To examine the joint and main effect of item type and proficiency on the test-takers' performance in the MSRT reading subtest, a Mixed ANOVA was conducted on SPSS. The prerequisite analyses disclosed no violation of the equality of covariances between the subgroups of proficiency (Sig. = 0.874). Similarly, Levene's test results supported the homogeneity of the four item types, including inference, paraphrase, information accuracy, and vocabulary. The relevant descriptive information for the MSRT reading item types is given in Table 7 and displayed by Figure 2 as follows.

Table 7. *Descriptive Statistics of the Four Item Types of the MSRT Reading Module*

Item Type	Proficiency	Mean	Std. Deviation	N
Inference	Lower	.4286	.21963	23
	Higher	.4208	.21808	37
Paraphrase	Lower	.5138	.22202	23
	Higher	.5946	.18507	37
Info_Accuracy	Lower	.3043	.36116	23
	Higher	.3649	.40223	37
Vocabulary	Lower	.5022	.23812	23
	Higher	.6095	.25051	37

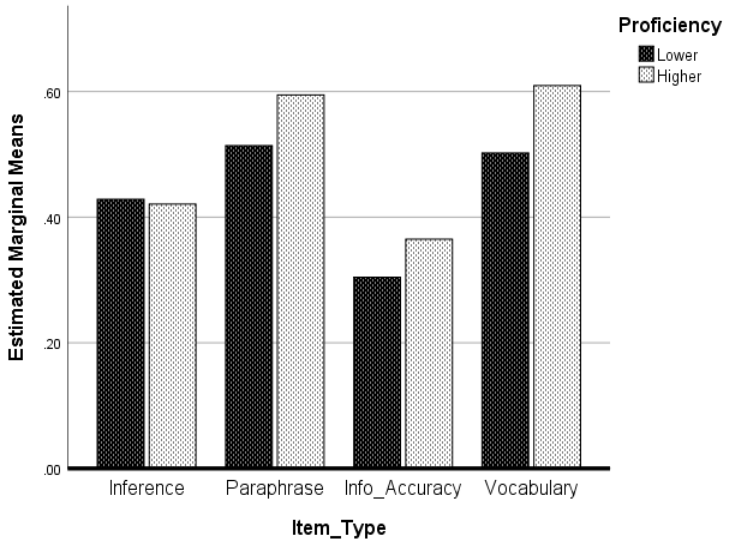

Figure 2. *Mean Scores of the Two Groups in Four Item Types of the MSRT Reading Module*

According to Table 7, the highest mean score (M = 0.60) represented the performance of the higher proficiency group in vocabulary, and the lowest mean score (M = 0.30) belonged to

the performance of the lower proficiency group in information accuracy items. Figure 2 delineates the comparison between the two groups of participants across four item types of the MSRT reading module. A quick review of the bars suggests the two groups performed almost equally (M = 0.42) under the inference questions, while their responses to the other item types were more distinct. In fact, in the paraphrase, information accuracy, and vocabulary items, the higher proficiency group had a comparatively better performance than the lower one.

Finally, the multivariate tests indicated that the interaction effect of item type and proficiency (based on Wilks' Lambda) was not significant (p-value = 0.292). Likewise, the tests of between-subjects' effects revealed no significant effect of proficiency as the observed p-value of 0.270 exceeded the common Alpha level. However, the effect of item type was found to be significant at 0.000, with a large effect size (partial eta squared = 0.41).

4. Discussion

The present study was an attempt to delve into the reliability of the reading module of MSRT, which is a well-known Iranian test of English proficiency. To this end, we set out to investigate how text topics, item types, and overall proficiency affect the examinees' reading comprehension.

4.1. Internal Consistency of MSRT Reading and its Internal Correlations

Based on the findings of this study, the MSRT reading module had quite a high level of internal consistency (0.86), indicating the homogeneity of the reading items in terms of measuring a single underlying trait. This finding, which may support the unidimensionality of the reading section of MSRT, is consistent with the findings reported by Rashvand Semiyari and Ahangari (2022) as well as those by Sahrai and Mamagani (2013) and Noori and Hosseini Zadeh (2017), who concluded the test carries an acceptable level of reliability. However, the mean of bivariate correlations amongst the four reading topics was 0.47, showing a moderate degree of go-togetherness amongst them. Specifically, the pairwise coefficients indicated that there was a large correlation between geology and history, geology and anthropology, and history and anthropology, which may prove these topics are more similar in measuring the same underlying construct.

As for the interrelationships amongst the four item types, a mean of 0.44 was observed, indicating they were moderately correlated. Of course, the correlation between inference and vocabulary item types was large, which implies inferring is probably more vocabulary-based. These results can partially corroborate Ghahraki et al. (2022), who validated the unidimensionality of the MSRT components. In summary, the findings reported through the present study disclosed although the reading subtest was found to carry a high degree of internal consistency, the correlations amongst the reading topics as well as the item types were moderate.

4.2. Effect of Topic and Item Type on MSRT Reading Performance

Taken together, the observed results showed that text topic played a significant role with a large effect size in the test-takers' reading achievement. The finding provides evidence that text topic can noticeably affect L2 reading performance, and thereby, endanger test reliability in the sense that it may be advantageous to some test takers due to their background topical knowledge.

This result, which agrees with several comparable findings (Ashrafzadeh et al., 2015; Khodi et al., 2024; Rashvand Semiyari & Ahangari, 2022), gives rise to the idea that topic bias may be unavoidable in any test of global language proficiency.

Like what was obtained for the influence of topic, item types in the MSRT reading subtest were found to trigger significant differences with a large effect size. Thus, the form and demand of the reading comprehension items can also affect the examinees' performance, leading to reliability problems. This recent finding, validated by Ahmadi Shirazi et al. (2019), reveals the fact that different item types challenge test takers differently as they tap into various dimensions of L2 reading comprehension. The variability seems to be rooted in item difficulty reported by scholars such as Ghahraki et al. (2022), who concluded that a considerable portion of the MSRT reading items is either very easy or very hard. It is a tenable argumentation that some item types in the reading section under scrutiny pose perceivably more difficulty for the test takers than some other items. Normally, this fluctuation has consequences for the reliability level of the MSRT test.

Considering the discussed findings, the various topics and item types in the MSRT reading subtest were found to be influencing the test-takers' reading performance significantly, which may produce uncertainties about unequal fluctuations across the examinees.

4.3. Influence of Overall Proficiency on MSRT Reading Achievement

The findings of this study signified that language proficiency, generally, did not have a significant effect on the MSRT test-takers' reading performance. This finding contrasts with those studies that position language proficiency as a powerful predictor of second language reading performance (e.g., Hill & Liu, 2012). Nevertheless, the significant interaction effect between topic and proficiency disclosed that the latter modified the test-takers' responses to different topics. The lack of a significant main effect for proficiency could probably reflect the difficulty of the test (Ghahraki et al., 2022), mitigating the possible distinguishing influence of proficiency. In fact, when the test items are too easy or too difficult, their discrimination will be reduced (Mehrens & Lehmann, 1991). In other words, the items cannot distinguish between high-ability and low-ability test takers (Ghahraki et al., 2022). As a piece of evidence in this case, our results exposed that proficiency failed to create a distinction between higher-ability and lower-ability test takers under the inference items. One possible reason could be the difficulty of such comprehension items as they need more reasoning to access the meaning implied by the text. To summarize, in three out of four contexts where the main and interaction effect of proficiency were measured, the grouping factor did not play a significant role in distinguishing the MSRT test-takers' reading performance.

Conclusion

The present study revealed that the MSRT reading module is quite a reliable measure of English reading comprehension in terms of its internal consistency, while the bivariate correlations implied an average degree of one-dimensionality amongst the text topics and item types. Indeed, the various text topics and item types emerged as significant factors, influencing the test-takers' reading performance. This finding might raise concerns about the fairness of such a national high-stakes test with long-lasting effects on the future of test takers. Therefore, the

MSRT designers may want to reconsider the role of topic and item type when developing the reading subtest to minimize the possible bias, which may otherwise jeopardize the test's fairness and equitability (Bachman & Palmer, 2010). The concluding remarks presented by the current study, therefore, have the potential to help with the improvement of the MSRT reading module and may contribute to better decision-makings towards the applicants with augmented stability, accuracy, and fairness.

Finally, it needs to be acknowledged that the present investigation involved a rather small sample size with only two proficiency levels, which could limit the generalizability strength of the reported findings. Further research with larger samples and more diverse proficiency groups is recommended to discover even more robust insights. Last but not least, a major limitation this study suffered from pertains to the fact that the Ministry of Science, Research, and Technology of Iran did not provide us with the real data produced by the past MSRT examinees. Certainly, sharing such data with researchers would allow for a more thorough investigation of the test, which can ultimately improve the reliability and validity of this famous national test of English proficiency.

References

Ahmadi Shirazi, M., Alavi, S. M., & Salarian, H. (2019). An investigation into item types and text types of reading comprehension section of Iranian Ph.D. entrance exams using G-theory. *Journal of Modern Research in English Language Studies,* 6(1), 1-29. https://doi.org/10.30479/jmrels.2019.10591.1326

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2019). *Introduction to research in education.* Cengage Learning.

Ashrafzadeh, A., Don, Z. M., & Meshkat, M. (2015). The effect of familiarity with content knowledge on Iranian medical students' performance in reading comprehension texts: A comparative study of medical and TEFL students. *Journal of Language Teaching and Research,* 6(3), 524-534. https://doi.org/10.17507/jltr.0603.07

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Best, J. W., & Khan, J. V. (2016). *Research in education*. United Kingdom: Pearson.

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Education ESL.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, 28(4), 523-539.https://doi.org/10.1016/S0346-251X(00)00036-1

Dewi, H., Damio, S., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education),* 9(1), 24-36. https://doi.org/10.21831/reid.v9i1.53514

Fallahian, E. & Tabatabaei, O. (2015). Construct validity of MSRT reading comprehension module in Iranian context. *English Language Teaching,* 8(9), 173-186. https://doi.org/10.5539/elt. v8n9p173

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage Publications Limited.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

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. https://doi.org/10.22034/ELT.2021.46325.2396

Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022*). A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Sage Publications, Inc.

Hashemi, A., & Daneshfar, S. (2018). A review of the IELTS test: Focus on validity, reliability, and washback. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, 3(1), 39-52. https://doi.org/10.21093/ijeltal. v3i1.123

Hill, Y. Z., & Liu, O. L. (2012). Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT® reading performance? *ETS Research Report Series*, 2012(2), 1-34. https://doi.org/10.1002/j.2333-8504.2012.tb02304.x

Karami, H. (2012). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. *Psychological Test and Assessment Modeling*, 54(3), 211-226.

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

Kiani, R., & Haghighi, M. (2006). The investigation of the TMU English proficiency test: Reliability related issues. *Quarterly Journal of Humanities,* 16(58), 55-73.

Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design* (3rd ed.). Taylor & Francis.

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

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.

Pallant, J. (2020). *SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS* (7th ed.). Routledge.

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. https://doi.org/10.30486/RELP.2021.1935588.1295

Sahrai, R. M., & Mamaghani, H. (2013). An assessment of reliability and validity of MSRT test. *Quarterly of Educational Measurement*, 3(10), 1-20.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497-516). Boston, MA: Pearson.