JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

Evaluating and Examining the Problems of Iran's Ministry of Higher Education and Universities' English Language Proficiency Tests and Stakeholders' Language Needs Analysis

Reza Pishghdam[*]
(corresponding author)
Professor in TEFL & Educational Psychology, Ferdowsi University of Mashhad, Mashhad, Iran
pishghadam@um.ac.ir

Shima Ebrahimi[**]
Assistant Professor of Persian Language Teaching, Ferdowsi University of Mashhad, Mashhad, Iran
Email: shimaebrahimi@um.ac.ir

Shaghayegh Shayesteh[***]
Assistant Professor of ELT, Ferdowsi University of Mashhad, Mashhad, Iran
Email: shayesteh@um.ac.ir

Sahar Tabatabaee Farani[****]
Ph.D. in TEFL, Ferdowsi University of Mashhad, Mashhad, Iran
Email: sahar.tabatabaee1983@gmail.com

Haniyeh Jajarmi[*****]
Assistant Professor of ELT, Bahar Institute of Higher Education, Mashhad, Iran
Email: hjajarmi@gmail.com

ABSTRACT

This study aims to explore English language proficiency tests of Iran's Ministry of Higher Education and universities by analyzing the needs of the participants (PhD candidates and graduates) and then propose a standard and uniform test. To this end, after thematizing the qualitative answers obtained through semi-structured interviews with 30 faculty members (19 females and 11 males) and 35 PhD candidates and graduates (13 females and 22 males) from different cities and majors in universities throughout Iran, a researcher-made questionnaire was developed. To validate the designed questionnaire, first 200 PhD graduates and candidates (117 males and 83 females), studying various majors in universities across Iran, answered the questionnaire (convenience sampling). Then 442 participants (223 males and 219 females) including PhD graduates (N= 57), PhD candidates (N= 320), and faculty members (N= 65) answered the validated questionnaire. Totally, 642 participants expressed their views on the problems of English language proficiency tests in terms of four criteria: validity, reliability, impact, and fairness of the test. The findings show that the participants are not satisfied with these tests and believe that these tests are designed based on the old versions of TOEFL and IELTS international tests and cultural and local considerations are overlooked. Finally, some suggestions are made to improve the quality of English language proficiency tests in Iran.

ARTICLE INFO
Article history:
Received:
24th, January, 2021
Accepted:
14th, February, 2021
Available online:
Winter 2021

Keywords:
examining the problems, English language proficiency tests, local-cultural components, Ministry of Higher Education

DOI: 10.22059/jflr.2021.317539.801

Pishghdam, Reza, Ebrahimi, Shima, Shayesteh, Shaghayegh, Tabatabaee Farani, Sahar, Jajarmi, Haniyeh (2021). Evaluating and Examining the Problems of Iran's Ministry of Higher Education and Universities' English Language Proficiency Tests and Stakeholders' Language Needs Analysis. *Journal of Foreign Language Research*, 10 (4), 686-705. DOI: 10.22059/jflr.2021.317539.801

[*] Reza Pishghdam is a professor in TEFL & educational psychology at Ferdowsi University of Mashhad. His research interests are neuro-psychological and socio-psychological aspects of language education, and social cognition.

[**] Shima Ebrahimi is an assistant professor of Persian language teaching at Ferdowsi University of Mashhad. Her areas of interests are educational psychology, sociology of language education, and teaching Persian to the speakers of other languages.

[***] Shaghayegh Shayesteh is an assistant professor of language education at Ferdowsi University of Mashhad. Her research interests are psychology and neuropsychology of language education.

[****] Sahar Tabatabaee Farani, a PhD holder in TEFL, has graduated from Ferdowsi University of Mashhad. Her research interest and focus are sociology, psychology, and neuropsychology of language education.

[*****] Haniyeh Jajarmi is an assistant professor of ELT. She has been engaged in interdisciplinary research within neuroscience, psychology of language, and language education. Her current research focuses on the neuro-psychology and socio-psychology of language education.

1. Introduction

A test is an instrument to measure a learner's knowledge and ability in a specific field and is considered to exert an effect on improving the quality of teaching (Hetman, Dreyfus, & Golan, 1990). In this vein, the reliability and validity of a test are of paramount importance (DeVon et al., 2007). A test with high validity and reliability indices can measure one's true level of ability in a field and pave the way for teaching and learning.

A test has an effect on teaching, which is called washback. In recent years, there have been many studies on the relationship between the washback effect and language learning (Roy Chan, 2020). A language proficiency test is a type of language test that reveals the level of the language learner's progress and his/her language abilities in all language skills regardless of a specific time and place (Abasi, 2015).

One of the important national tests in Iran is the English language proficiency test for PhD candidates that is held by the Ministry of Higher Education and some of the universities in Iran. As per the important international role of the English language, language proficiency tests are administered to assess graduate students' level of proficiency as a permit to continue their education at the PhD level. PhD candidates are supposed to read and write ESP (English for specific purposes) texts; that is, they should have the ability to write, read, and comprehend articles, dissertations, and texts in English; they should even be able to present the results of their scientific studies in international conferences. In most cases, the four basic competencies (i.e., listening, speaking, reading, and writing) are evaluated in proficiency tests to show how competent the test takers are in terms of language skills. In such tests, the students' overall language ability indicates their linguistic knowledge, familiarity with language components, and correct use of language forms (Farhady, Ja'farpur, & Birjandi, 2007). It seems that language proficiency tests are suitable choices to determine students' level of language skills.

Emphasizing the four language skills, the national English language proficiency tests try to be similar to the international ones such as IELTS and TOEFL. However, most of them cover vocabulary, grammar, listening comprehension, and reading comprehension and disregard speaking and writing. As a matter of fact, these tests do not benefit from a systematic look at the language and stick to rote learning by compartmentalizing language skills. Hence, designing standard tests consisting of all language skills and tailored to the test takers' field of study seems necessary. Therefore, the present study sets out to examine and evaluate the current English language proficiency tests in Iran and present a series of suggestions to design standard test items.

2. Literature Review

Council of Europe (2001) defines language proficiency based on The Common European Framework of Reference (CEFR). As per this definition, competence means all skills and abilities that someone may need to communicate, and it is divided into general and communicative

parts. The general part is the language-independent knowledge and skill of learning, and the communicative part includes language use and is evaluated through linguistic, sociolinguistic, and pragmatic competencies. CEFR describes six levels of language competence (A1, A2, B1, B2, C1, C2) ranging from elementary to advanced levels (Harsch, 2017).

There have been different views about language proficiency during the last decades. As a case in point, during the 1960s, Lado (1961) and Carroll (1961) introduced the Skill-Component Model of language proficiency with different language skills (listening, speaking, reading, and writing) and language components (grammar, vocabulary, phonology/graphology). This view was based on structuralism and led to discrete-point tests to evaluate language proficiency (Motallebzadeh & Baghaee Moghaddam, 2011). The Skill-Component Model had some weaknesses, as well. For instance, it was not clear if skills were different from components or they were closely related to each other. Lack of a situational context of language use was considered as another limitation of the Skill-Component Model (Backman, 1990).

Criticizing Backman's (1990) model of language proficiency, Roever and McNamara (2006) discussed pragmatic, cultural, and test bias aspects and emphasized the social aspect of a language test. While the pragmatic aspect centers around evaluating language use in real-life conditions, cultural and test bias aspects focus on those test takers who take advantage of the test content unfairly. Furthermore, Roever and McNamara (2006) focused on the effects of test bias and cultural issues of a test and highlighted the need for more research studies on the social effects of language tests.

Bachman and Palmer (2010) assessed language evaluation systems in a broader sense. They believed that,

> the AUA consists of a set of claims that specify the conceptual links between a test taker's performance on an assessment, an assessment record, which is the score or qualitative description we obtain from the assessment, an interpretation about the ability we want to assess, the decisions that are to be made, and the consequences of using the assessment and of the decisions that are made. (p. 30)

In order to emphasize the usefulness of assessment and its interpretations, the AUA presents four claims: 1. Test consequences and decisions are useful; 2. Decisions are made based on the values and equitability of the target society; 3. The interpretations of language ability include concepts such as meaningfulness, impartiality, generalizability, relevance, and sufficiency; 4. Assessment scores show consistency (Clark-Gareca, 2010). In the third claim, meaningfulness means that the assessed construct represents the same meaning for all test takers; impartiality implies that the content and the construct of the test are not biased toward a specific group; generalizability argues that the results of the test can be generalized to other contexts; relevance deals with the fact that whether the test assesses the intended construct; and, finally, sufficiency asks whether the test

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

provides sufficient information for decision making (Bachman & Palmer, 2010). Based on different definitions of language proficiency and the purposes of each definition, several language proficiency tests (for different languages) have been designed. Table 1 includes English language proficiency tests designed and applied in Iran.

Table 1. The Formats of Iran's English Language Proficiency Tests

	Test Holder	Test Title	Test Type	No. of Items	The Skill Assessed					
					Vocabulary	Grammar	Listening	Reading	Writing	Speaking
1	Ministry of Higher Education	MSRT	Multiple-Choice	100		√	√	√		
2	National Organization of Educational Testing	TOLIMO	Multiple-Choice & Essay-Type	141		√	√	√	√	
3	Ministry of Health and Medical Education	MHLE	Multiple-Choice	100	√	√	√	√		
4	Islamic Azad University	EPT	Multiple-Choice	100	√	√		√		
5	Tehran University	UTEPT	Multiple-Choice	100	√	√		√		
6	Tarbiat Modares University	TMUE	Multiple-Choice	100	√	√		√		
7	Ferdowsi University of Mashhad	TELP	Multiple-Choice	100		√	√	√		
8	Isfahan University	UIEPT	Multiple-Choice	80	√	√		√		
9	Shahid Chamran University of Ahvaz	SCU	Multiple-Choice	100	√	√		√		
10	Payame Noor University	ETPNU	Multiple-Choice	100		√	√	√		
11	Kharazmi University	KELT	Multiple-Choice	100	√	√		√		
12	Urmia University	UUEPT	Multiple-Choice	100		√	√	√		
13	Shiraz University of Technology	SELT	Multiple-Choice	100	√	√	√	√		
14	Razi University	RULPT	Multiple-Choice	100	√	√		√		
15	University of Shahrekord	SKELT	Multiple-Choice	100	√	√	√	√		

Except for TOLIMO that includes one essay-type item in addition to multiple-choice items, other language proficiency tests are made up of multiple-choice items and include up to 100 items. Based on the information about language proficiency tests in Iran and what is presented in Table 1, these tests consider writing skill as the most complicated one that challenges the test takers (Nunan, 1989), so they ignore this skill. However, it should be noted that the level of thinking and understanding can grow by writing (Bazerman, Little, & Bethel, 2005), and assessing this skill can demonstrate test takers' comprehension and thinking in a better way. Pishghadam and Ebrahimi (2019) believe that writing skill plays a significant role in improving linguistic abilities, level of understanding, communicating with others, and expressing emotions; yet, this skill has been stigmatized, making Iranian students suffer from weaknesses in this field.

As Table 1 demonstrates, only TOLIMO, administered by the National Organization of Educational It should be noted that the writing skill in TOLIMO has a separate score that is added to the overall score. TOLIMO is similar to IELTS and TOEFL in assessing the writing skill and reporting its score. The speaking skill is ignored in all tests. All centers (15 centers) have focused on the reading skill and grammar sections. The vocabulary (10 centers) and the listening skill (8 centers) sections come as the next priorities. All of these centers have claimed that the test items are comparable to those of TOEFL and IELTS. However, it should be regarded that the international tests of TOEFL and IELTS deal with all four skills separately, and each skill takes a separate score (Terry & Wilson, 2004).

Language proficiency tests in Iran do not boost the test takers' level of English language; that is, when they prepare themselves for the test, they just stick to rote learning for the sake of the score. Speaking and writing skills are also important for PhD candidates, but not sufficient attention has been paid to them. Furthermore, these tests suffer from a lack of validity and may have negative effects on the target society. Hence, examining the present tests and suggesting a validated alternative adapted to the local culture is worthy of attention.

3. Methodology

Participants

The present mixed-methods study employed qualitative interviews and quantitative data analysis to scrutinize the English language proficiency tests designed by the Ministry of Higher Education and some of the universities of Iran. In the qualitative phase, we interviewed 30 faculty members (19 females and 11 males) and 35 PhD holders and PhD candidates (13 females and 22 males) from different cities and university majors who have participated in several English language proficiency tests. The interviews continued until data saturation was reached.

A number of 642 participants took part in the quantitative phase. In the first step, 200 PhD holders and PhD candidates (117 males, 83 females) from different majors and universities of Iran filled out the designed questionnaire. Afterward, the designed questionnaire was substantiated using 442

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

participants (223 males, 219 females), including PhD holders (N = 57), PhD candidates (N = 320), and faculty members (N = 65).

Instrumentation

Semi-structured interviews were conducted in the qualitative phase. The questions are listed below:

1. Which English language proficiency tests have you already taken?
2. Do you think English language proficiency tests can assess language skills correctly?
3. Are you satisfied with the content of the tests?
4. Do you think the test administration conditions (i.e., setting and time) are suitable?
5. How can these tests impact your academic life?
6. What are the positive and the negative aspects of these tests?
7. What is your suggestion to improve the quality of these tests?

Based on the data obtained from the interviews, a 25-item questionnaire was designed in the quantitative phase. Confirmatory Factor Analysis (CFA) was applied to validate the designed questionnaire (Appendix 1).

Data Collection

To collect the data, the participants were first invited to share their ideas about English language proficiency tests. The time period for each interview was almost 45 minutes, and the data collection procedure lasted for two months (January 2020 to March 2020). Prior to the interviews, the interviewers explained the concept of language proficiency to the participants who did not know it. After collecting the required data, the researchers transcribed the recorded interviews and categorized the extracted themes. Guba and Lincoln (1994) considered the four criteria of credibility, dependability, confirmability, and transferability to verify the validity and the reliability of the qualitative data in a study (Mohsenpoor, 2011). The qualitative data analysis was further checked by two professors of TEFL (Teaching English as a Foreign Language), an assistant professor of Persian Language Teaching, an assistant professor of TEFL, and two PhD holders of TEFL. After data was reached, classification of the data was done, and the face validity of the content was examined. The extracted data were analyzed by the researchers. Subsequently, the most frequent responses were named as the main categories. These categories included validity, reliability, test effect, and test fairness, each of which had some subcategories. For each category, six items were designed. The 24-item questionnaire was applied to collect the data for the quantitative phase of the study (Appendix 1). Table 2 presents the major and minor themes.

Table 2. The Themes Extracted From the Interviews

Row	Subcategories	
1	Test Validity	Assessing participant's language ability
		Conforming the local and cultural norms
		Assessing all language skills
		Assessing language totality
2	Test Reliability	Test administration conditions
		Test content
3	Test Effect	Language learning
		Mind and body
		Expenditures
		Professional career
		Participants' needs analysis
4	Test Fairness	Handling complaints
		Bias
		Criterion score
		Exam preparation courses
		Fairness
		Availability of sample items

After categorizing the responses, a questionnaire was prepared by the researchers and checked by the experts in the field to verify its content and face validity. To substantiate the construct validity of the questionnaire, we administered it to 442 participants (223 males, 219 females), including 57 PhD holders, 320 PhD candidates, and 65 faculty members on Google Forms. The data were analyzed using repeated-measures analysis of variance (ANOVA) employing the SPSS software (version 21).

4. Results

Qualitative Data Analysis

After examining the participants' responses to the interview questions, we extracted four main themes: validity, reliability, test effect, and test fairness.

Validity

Validity deals with the extent to which an instrument measures what it claims to measure. According to the responses presented in Table 3, the participants believe that English language proficiency tests cannot measure language skills

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

correctly and do not benefit from a high level of validity.

Table 3. The Subcategories of Test Validity

	Subcategory	
1	Unsuitability of the tests to measure the participant's language proficiency (93%)	Ignoring the language proficiency level (91%)
2	Lack of conformity to local and cultural norms (94%)	Copying international tests (100%)
3	Failure to assess all language skills (89%)	Ignoring the speaking skill
		Ignoring the writing skill
4	Failure to assess language ability (100%)	

Reliability

Reliability indicates the consistency of results upon multiple administrations. Based on the responses of the participants, the subcategories of reliability and the participants' opinions about them are presented in Table 4.

Table 4. The Subcategories of Test Reliability

	Subcategory		
1	Test administration condition is not suitable (37%)	Test time is not suitable (79%)	Test administration time is not suitable (42%)
			Test duration is not suitable (58%)
		Test setting is not suitable (56%)	
2	Test content	The number of items is not suitable (21%)	
		The test is biased towards a specific field of study (84%)	Test item preparation is biased (47%)
		Level of difficulty	Unprofessional test developers (28%)
			Lack of a specific framework (67%)
			Difficult reading passages (94%)
			Difficult grammar items (43%)
			Mistake in the test items (57%)
		The options	Mistake in the options (64%)
			Random selection of the options (75%)

Impact

In addition to the validity and reliability of the test, several responses were related to the effect that the tests have on the participants' academic and professional life. In this regard, the authors divided the effects of the test into the following subcategories.

Table 5. The Subcategories of Impact

	Subcategories	
1	The test does not affect language learning (83%)	
2	The test negatively affects the mind and body of the participants (91%)	Creating stress and anxiety
		Causing palpitations, headaches, stomach aches, etc.
3	The cost of the test is not appropriate (69%)	Having a high cost
		Expensive test preparation classes
		High cost of travel
4	The test does not affect the participants' profession (79%)	
5	The test does not consider the real needs of the participants (100%)	Disregarding essay writing skills
		Not paying attention to the public speaking skills

Fairness

Subsequently, one of the concerns of most of the interviewees was the fairness of the test. Based on the responses, the themes that fall into the category of test fairness are summarized in Table 6.

Table 6. The Subcategories of Test Fairness

	Subcategories	
1	The way the objections to the test are handled is not fair (21%)	To a wrong question or option
		To the test results
		To the quotas
2	The test is biased towards some particular fields of study (63%)	
3	The criterion score is not fair. (73%)	Test score
		The acceptable scores for universities
4	Attending test preparation classes is only possible for some of the participants, and this is not fair (92%)	
5	The test items are not designed fairly (72%)	Regarding different majors
		Regarding the difficulty level of the test
6	Sample test questions are unavailable (65%)	

Ninety-one percent of the interviewees agreed that the national tests were not designed under Iranian cultural and local standards and were merely imitations of the international TOEFL and IELTS tests (paper-based) and said that: "*It is natural for a test, that tries to be a copy of international tests, not to meet the cultural needs of Iranians*"; "*Of course, the test designers should be given the right not to design tests based on local cultural needs. Besides the fact that cultural needs are not so important in measuring language knowledge, they try to make the test more similar to the international tests to seem more standard"*; "*Universities have been trying*

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

to design a separate test for themselves for years, but when we put all these tests together, we still see that they are the simplified imitations of the TOEFL and IELTS tests"; "*Unfortunately, copying of international exams has been done deficiently and has reduced the quality of these exams*"; "*I think the questions should become localized. Since the comprehension questions are related to an English subject, they do not measure real comprehension of the content*". However, some candidates considered these similarities as an advantage and expressed their opinion as follows: "*I think these tests should be designed following international standards and it is very good that they are similar to international tests. It is possible to take a national test similar to the international ones at a much lower cost and achieve the desired goal, which is to get a certificate*". According to most of the participants, these tests are based on excerpts from the old international TOEFL and IELTS tests, and even though tests often have a cultural load, cultural and local considerations are usually ignored.

Quantitative Data Analysis

Descriptive statistics for the designed questionnaire and the participants' age are shown in Table 7. As can be seen, the lowest mean value is related to "impact" (M = 2.41), and the highest value is related to "reliability" (M = 2.95).

Table 7. Descriptive Statistics for the Designed Questionnaire and the Participants' Age

Construct	N	Minimum	Maximum	Mean	SD
Validity	200	1.00	5.00	2.50	0.92
Reliability	200	1.00	5.00	2.95	0.85
Impact	200	1.17	5.00	2.41	0.87
Fairness	200	1.00	4.67	2.80	0.67
Age	200	23	58	34.37	6.19

Prior to substantiating the construct validity of the questionnaire, the normality of the data was examined. As can be seen in Table 8, the values of Skewness and Kurtosis were in the range of 2 and –2, indicating the normal distribution of the data.

Table 8. Descriptive Statistics for the Normality of the Data

	Skewness	Kurtosis
Questionnaire	0.29	–0.04

The Harman one-factor test was performed to avoid the common method variance (CMV) error. The results of the Harman test showed that the first factor accounted for 25% of the total variance. As a result, it can be concluded that the error of the common method is not problematic in this study. After the Harman test, the construct validity of the questionnaire was assessed through CFA. Figure 1 shows the CFA model for the designed questionnaire. As can be seen, the questionnaire consists of 4 subscales (validity, reliability, impact, and fairness), each of which is measured by six items.

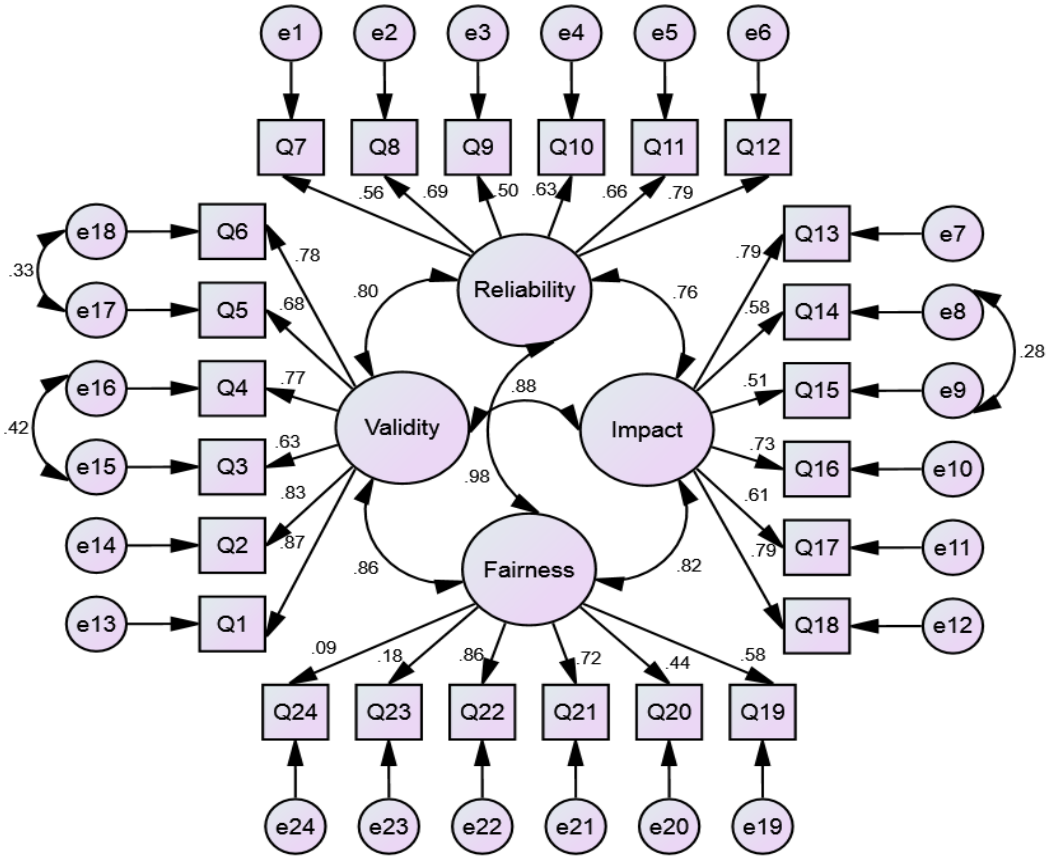

Figure 1. The CFA Model for the Designed Questionnaire

Table 9 presents the goodness of fit indices. The reported indices (the chi-square to the degree of freedom ratio ($\chi 2$/df), the comparative fit index (CFI), Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean squared residual (SRMR)) approve the model's goodness of fit. Based on Ulman (2001), to maintain model fit, the $\chi 2$/df value needs to be less than 3. Moreover, based on Browne and Cudeck (1993), the CFI and TLI indices should be above 0.90, and the RMSEA should be less than 0.8.

Table 9. The Goodness of Fit Indices

$\chi 2$/df	df	CFI	TLI	RMSEA	SRMR
1.95	243	0.91	0.90	0.06	0.05

As can be seen, the model has a good fit without removing any items. Therefore, the validity of the designed questionnaire has been confirmed. After validation, the reliability of the questionnaire and its four constructs was examined using Cronbach's alpha (Table 10). The results showed that all values were statistically desirable (above 0.70).

Table 10. The Reliability of the Questionnaire and its Four Constructs

	Number of Questions	Cronbach's Alpha
Total	24	0.93
Construct 1 (Validity)	6	0.90
Construct 2 (Reliability)	6	0.81
Construct 3 (Impact)	6	0.81
Construct 4 (Fairness)	6	0.70

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

The normality of the collected data was examined for further analysis. As Table 11 shows, the Skewness and Kurtosis values of "validity", "reliability", "impact", and "Fairness" were in the range of –2 and 2, indicating the normal distribution of the data.

Table 11. Normality Test Results for the Four Constructs of the Questionnaire

Constructs	Skewness	Kurtosis
Validity	0.61	-0.13
Reliability	-0.11	-0.51
Impact	0.71	0.24
Fairness	-0.15	0.00

To examine if there are any significant differences among the mean values of the four constructs of the questionnaire, repeated measures analysis of variance (ANOVA) was used. As Table 12 shows, there is a significant difference among the mean values of validity (M = 2.48), reliability (M = 2.88), impact (M = 2.40), and fairness (M = 2.75) [F (3, 1323) = 97.21, p = 0.00, $\eta_p^2 = 0.18$, $1-\beta = 1.00$]. In this analysis, the value of the partial Eta squared shows a large effect size, and the observed power indicates that the sample size is sufficient.

Table 12. Tests of Within-Subjects Effects for Comparing the Mean Differences among the Questionnaire's Four Constructs

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Observed Power
Constructs	Sphericity Assumed	68.24	3	22.74	97.21	0.00	0.18	1.00
	Greenhouse-Geisser	68.24	2.75	24.80	97.21	0.00	0.18	1.00
	Huynh-Feldt	68.24	2.77	24.63	97.21	0.00	0.18	1.00
	Lower-bound	68.24	1.00	68.24	97.21	0.00	0.18	1.00

Since the mean values of the four constructs of validity, reliability, impact, and fairness were significantly different, pairwise comparisons were performed to determine where the differences occurred (Table 13).

Table 13. Pairwise Comparisons of the Four Constructs of the Questionnaire

Construct I	Construct J	Mean Differences (I-J)	Std. Error	Sig. [a]
1	2	–0.40*	0.35	0.00
	3	0.08*	0.30	0.04
	4	–0.26*	0.33	0.00
2	1	0.40*	0.35	0.00
	3	0.48*	0.36	0.00
	4	0.13*	0.29	0.00

Construct I	Construct J	Mean Differences (I-J)	Std. Error	Sig. [a]
3	1	–0.08*	0.30	0.04
	2	–0.48*	0.36	0.00
	4	–0.34*	0.31	0.00
4	1	0.26*	0.33	0.00
	2	–0.13*	0.29	0.00
	3	0.34*	0.31	0.00

[a]. Adjustment for multiple comparisons: Bonferroni

*. The mean difference is significant at the .05 level.

To sum up, the results of the repeated-measures analysis of variance and pairwise comparisons of the four constructs of validity, reliability, impact, and fairness indicated that:

(2.88) Reliability (2.75) < Fairness (2.48) < Validity (2.40) < Impact

5. Discussion and Conclusion

Tests are tools to measure participants' academic achievement in educational settings. In the process of designing a test, some principles must be considered to achieve the educational goals. In this regard, English language proficiency tests are important measurement tools that normally assess language competence, communicative competence, and the ability to apply knowledge of the language to real-life situations. The results of these tests are crucial in the field of education since they are a prerequisite for the comprehensive doctoral exam for various disciplines. Most proficiency tests are designed at the advanced level. At this level, the participants are supposed to have a sufficient knowledge of vocabulary, expressions, grammar, and the four skills of listening comprehension, speaking, reading comprehension, and writing at the native speakers' level of language proficiency. Given these factors, the importance of English language proficiency tests cannot be ignored.

An investigation of the participants' views and the findings of the study showed that in Iran, English language proficiency tests do not enjoy the desired validity. Given that validity refers to the accuracy of the results and shows the extent to which the test correctly measures what it intended to measure (Mohammad Beigi, Mohammad Salehi, & Gol, 1393), it cannot be expected that the language proficiency level of the participants was assessed correctly if the test was not valid. Therefore, the impact of these tests is also questionable.

According to Kachru's (1986) World Englishes theory, the acceptance and revival of local languages are of high significance in education. Based on this concept, different types of Englishes emerge through localization (adopting English speakers' culture with the

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

native culture), linking (combining English speakers' culture with the native culture), and acculturation (accepting English speakers' culture) (Tupas, 2004). It seems that all three can be considered in designing English language proficiency tests. Therefore, to comply with the requirements of the integrated quality management standards of English language proficiency tests, it is necessary that the Ministry of Education design a standard test following the cultural and local needs of the test takers and inform universities of how to implement it.

Based on the research findings, it seems that the universities in charge of these exams merely aim to take the test, and the students are obliged to take part in the test and get the required score. However, it should be noted that requesting a certificate from students without providing adequate infrastructure is an underestimation of the issue. It is clear that participants will eventually study in a short and intensive course of time simply to get a certificate in a short time and will forget most of the material after the test. Since learning is purely parrot-like and superficial, it will only remain in the participant's memory for a limited time. As mentioned earlier, a standard test enjoys the psychometric properties of validity and reliability. Besides, the content of the questions and options and the absence of mistakes in them are essential. However, the research findings showed that in some cases, the questions might measure insignificant information so that the participants may get confused as a result of the simplicity of the item and choose the wrong option. In some other cases, on the contrary, the questions may be so difficult that the participant may randomly decide to choose one of the options.

In general, given that evaluation has an important role in the educational system, in addition to static methods that believe that the participants' scores represent their full mental capacity, information, and skills (Ghafourian & Ashouri, 1396), dynamic assessment is also very substantial. This type of assessment considers the two processes of teaching and assessment as a combined activity to find out the abilities of learners and help them promote these abilities (Poehner, 2008); therefore, the developing skills of the candidates are better demonstrated. This provides a complete picture of the candidates' hidden abilities and differences, and taking into account the individual characteristics of the students, it is a good alternative to static assessment. In such situations, dynamic assessment is employed to achieve a better understanding of students' strengths and weaknesses and help them improve their language skills (Poehner, 2009).

It is evident that the level of preparation of the candidates and the conditions of the test are also very important and affect the interpretation of the test results. Therefore, these conditions are to be as standard as possible and the same for all candidates (Colemen, 1998). Mohammadi Roozbehani (1385) believes that the candidates' satisfaction with the test depends on the comfort they experience during the test, environmental conditions, equipment, being notified of the instructions before and during the test, difficulty level of the test, and their prior preparation. These factors are fundamental to the standardization of the test and fall into the category of test reliability.

As the findings show, while the candidates are relatively satisfied with the conditions of the test, they are dissatisfied with the content of the tests and their difficulty level. In fact, contrary to the claims of the test administrators, who try to make the test similar to the international ones by emphasizing the four language skills, many of these tests consider only a limited part of the language knowledge (often vocabulary, grammar, and reading comprehension) while they overlook the more important skills, like speaking and writing. It is clear that in such situations, the candidates only prepare themselves for grammar, vocabulary, and reading comprehension skills and do not pay much attention to other skills. Accordingly, it seems that English language proficiency tests and their sub-sections, such as listening comprehension, vocabulary, grammar, reading comprehension, etc., should not be separated from each other. In other words, the approach to testing should be academic, authentic, and integrated, and the sections should not be fragmented. Therefore, it is suggested that reading and translating, listening and speaking, and listening and writing be evaluated together, and if vocabulary and grammar are to be considered, they might be embedded in these sections.

In general, based on the results of the current proficiency tests, it is observed that these tests not only do not have a positive effect on the learning and performance of the candidates but also have a negative washback effect. In designing these tests, it is crucial to consider what the main goals of holding the tests are and who the administrators, participants, and stakeholders are. Therefore, benefitting from the results of this study, we can suggest a standard test with a systematic approach in line with the local-cultural components that will change the process of the aforementioned tests. Based on the findings of the current study, the following suggestions are made to improve the quality of English language proficiency tests:

1. It is better to hold a professional look at testing and to establish an assessment center in each university to monitor and standardize the tests.
2. If there is a single assessment center for monitoring and standardization of tests, English language proficiency tests held in universities be supervised by the Ministry of Higher Education.
3. It is suggested that the test be held in clusters for different fields of study. It is evident that the nature of each field of study is different from the other, and it is not possible to design a single test in which there is no bias.
4. Efforts should be made to assess all language skills in English language proficiency tests and to consider speaking and writing skills along with other skills.
5. Efforts should be made to design the tests in English based on the philosophy of "World Englishes" and Iranian culture.
6. Tests in only the multiple-choice format that is memory-based should be avoided.
7. The fragmentation of language to grammar and vocabulary is not scientific and is not recommended.

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 686 to 705

8. It is suggested that the Ministry of Higher Education tests of language proficiency be held online and on virtual platforms in all universities.
9. It should be possible to hold the test online.
10. To establish educational fairness, sample tests should be provided for the candidates on a website.
11. As the designing of standardized tests is costly, it is suggested that ministries and universities provide subsidies for the tests and do not just rely on the costs they receive from the candidates.
12. Efforts should be made to design tests based on the dynamic approach to testing and the school of "Test for Learning".
13. The test questions bank should be constantly monitored.
14. It is suggested that candidates be allowed to obtain 30% of the score by participating in standard university language classes and 70% by taking the test.
15. The number of test administration centers should be increased to reduce trips to the test centers and approach educational fairness.
16. The sections of the test are suggested to include a) reading comprehension and translation; b) listening and speaking; c) listening and writing.

In general, the present study suggests the administration of the test in the following ways:

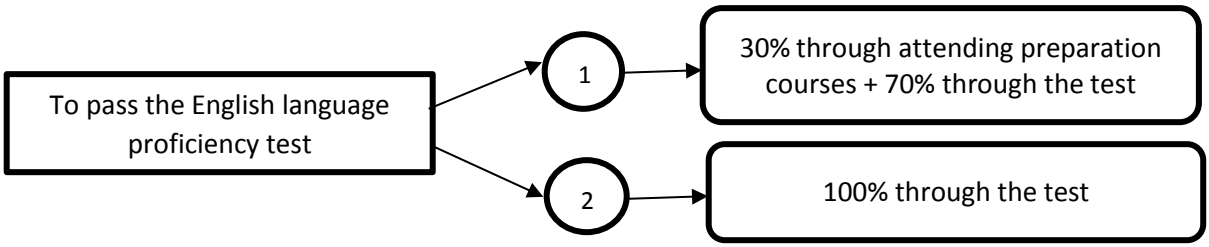

Figure 2. The Proposed Ways of Holding English Language Proficiency Tests

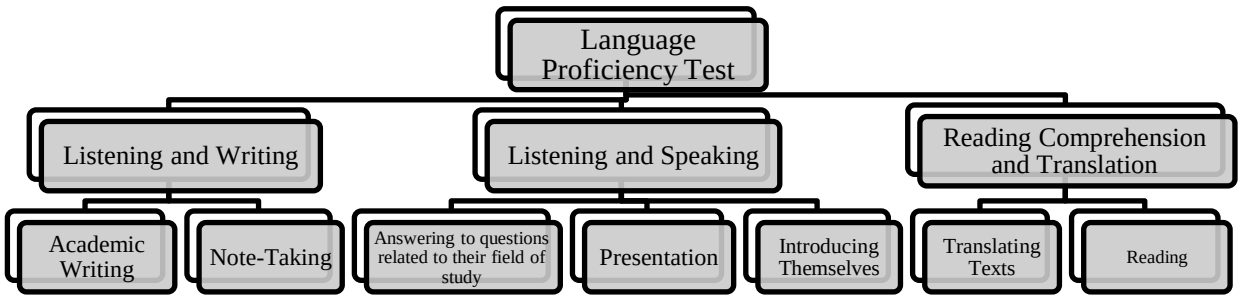

Figure 3. The Proposed Format for English Language Proficiency Tests Based on the Four Skills

Due to the importance of designing the tests based on the candidates' needs, researchers should consider the cultural norms of Iranian society in compiling test questions. On the other hand, English language proficiency tests for specific purposes can be designed which comply with the needs of the candidates in each field. In some cases, the test can be taken orally for some disciplines. Therefore, accurate needs analysis of the candidates based on their field of study can be one of the considerable topics for future research.

References

In Persian

Abbasi, Z. (March, 2015). Tahiye va tadvin-e azmoon-e basandegi-e zaban-e farsi [Preparation and compilation of Persian language proficiency tests]. *Paper presented at the First Conference on Teaching Persian Language*. Tarbiat Modares University, Tehran.

Ghafourian, M., & Ashouri, M. (1396). Daneshamoozan-e dir amooz: arzyabi-e pooya, vizhegi ha, shenasaee, shive hay-e tadris va behbood-e zarfiat-e yadgiri [Inapt learners: Dynamic assessment, characteristics, identification, teaching methods, and improving learning capacity]. *Exceptional Education, 17*(4), 57-64.

Mohammad Beigi, A., Mohammad Salehi, N., & Gol, M. A. (1393). Ravaee va payaee-e abzarha va ravesh hay-e mokhtalef-e andaze giri-e anha dar pazhuhesh hay-e karbordi dar salamat [Validity and reliability of various instruments and their measurement methods in applied research in health]. *Journal of Rafsanjan University of Medical Sciences, 13*(12), 1153-1169.

Mohammadi Roozbehani, K. (1385). Moghadameie bar standard sazi-e ejray-e azmoon ha: barrasi-e sharayet-e bargozari-e azmoon hay-e sarasari-e vorood be daneshgah ha az rah-e sanjesh-e rezayatmandi-e sherkat konandegan [Introduction to standardization of exams: Investigating the conditions of holding national entrance exams to universities by measuring the satisfaction of participants]. *Research and Planning in Higher Education, 3*(41), 1-12.

Mohsenpour, M. (1390). Arzyabi-e dade yay-e keifi [Evaluation of qualitative data]. *Bayhaq, 16*(2), 50-55.

Pishghadam, R., & Ebrahimi, S. (1397). Nesbiat-e hesi va tasir-e an bar maharate neveshtari zaban amoozan-e gheir-e farsi zaban [Sensory relativism and its effect on the writing skills of non-Persian language learners]. *Journal of Language-Related Research, 9*(6), 213-240.

In English

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

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

Bazerman, C., Little, J., & Bethel, L. (2005). *Reference guide to writing across the curriculum.* US, South Carolina, Anderson: Parlor Press LLC.

Carroll, J. B. (1961). The nature of data, or how to choose a correlation coefficient. *Psychometrika, 26*(4), 347-372.

Clark-Gareca, B. (2010). Language assessment in practice. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 10*(2), 41-46.

Coleman, A. L. (1998). Excellence and equity in education: High standards for high-stakes tests. *Virginia Journal of Social & the Law, 6*(10), 81-114.

DeVon, H. A., Block, M. E., Moyle Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D.J., et al. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155-64.

Farhady, H., Jafarpur, A., & Birjandi, P. (2007). *Testing language skills: From theory to practice*. Tehran: SAMT.

Harsch, C. (2017). Proficiency. *ELT Journal, 71*(2), 250-253.

Herman, J., Dreyfus, J., & Golan, Sh. (1990). *The effects of testing on teaching and learning*. Los Angeles, CA, National Center for Research on Evaluation, Standards, and Student Testing.

Kachru, B. B. (1986). *The alchemy of English: The spread functions and models of non-native Englishes*. Oxford, UK: Pergamon.

Lado, R. (1961). *Language testing*. New York, NY: McGraw-Hill.

Motallebzadeh, K., & Baghaee Moghaddam, P. (2011). Models of language proficiency: A reflection on the construct of language ability. *Iranian Journal of Language Testing, 1*(1), 42-48.

Nunan, D. (1989). *Second language teaching & learning*. Boston: Heinle & Heinle.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin: Springer.

Poehner, M. E. (2009). Group dynamic assessment: Mediation for the L2 classroom. *TESOL Quarterly, 43*(3), 471-491.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics, 16*(2), 242-258.

Roy Chan, K. L. (2020). Washback in education: A critical review and its implications for

language teachers. *Journal of Foreign Language Education and Technology, 5*(1), 108-124.

Terry, M., & Wilson, J. (2004). *Focus on academic skills for IELTS*. London, England: Pearson Education Limited.

Tupas, T. R. F. (2004). The politics of Philippine English: Neocolonialism, global politics and the problem of postcolonialism. *World Englishes, 23*(1), 47-58.

Yin, R. K. (2010). *Qualitative research from start to finish*. New York, NY: The Guilford Press.

Appendix

Sample Questions of the Questionnaire for Examining the Problems of English Language Proficiency Tests

1. The test correctly measures the participant's language proficiency.
2. The test measures the four main language skills.
3. The test is held in a standard condition.
4. Taking this test does not impose much cost.
5. Passing the test has a positive impact on my career.
6. The test is not memory-based.