IMAM KHOMEINI INTERNATIONAL UNIVERSITY

Journal of Modern Research in English Language Studies

Print ISSN: 2676-5357
Online ISSN: 2676-5985

IELTS and TOEFL Applicants’ Self-Assessment and Actual Test Performance: A Mixed-Methods Study of Relationship and Variation

Reza Rezvani[1*], Farshad Khadivikia[2]

[1*](Corresponding author) Associate Professor, Department of English Language and Literature, Yasouj University, Yasouj, Iran. *rezvanireza@gmail.com*

[2]MA in TEFL, Department of English Language and Literature, Yasouj University, Yasouj, Iran. *farshadkhadivikia@gmail.com*

Article info

Article type: Research article

Received: 2024/05/14

Accepted: 2024/08/07

Abstract

Recent research on self-assessment (SA) has primarily dealt with its relationship with students’ scores. However, few studies have systematically explored SA in language proficiency tests. This mixed-methods study examined differences between IELTS and TOEFL applicants’ self-assessments and their actual test scores. It also explored the sources of variations between these two assessments. The study sample included 81 IELTS (n= 51) and TOEFL (n=30) participants. Data collection involved the applicants' self-assessments, their test scores, and semi-structured interviews. Quantitative data were analysed using correlation and regression analyses, while qualitative data were examined through thematic content analysis. The statistical analyses revealed moderate to moderately high correlations between self-assessed and actual test scores. Self-assessments in speaking, reading and writing for both IELTS and TOEFL showed strong correlations with test scores. However, self-assessments in listening were only moderately correlated with actual test scores of both IELTS and TOEFL. In addition, regression analyses indicated that self-assessments in reading, speaking and writing for IELTS, as well as reading and speaking for TOEFL, were significant predictors of test scores. On the other hand, self-assessments in listening for both IELTS and TOEFL, as well as TOEFL writing, were poor predictors of actual test scores. Furthermore, qualitative data analyses highlighted the influence of factors such as experience, psychological aspects, linguistic abilities, background knowledge and feedback in explaining the variations between self-assessments and actual test performance. In conclusion, the paper discusses the findings and implications of the study in the context of language proficiency testing

Keywords: high-stakes assessments, IELTS, Iranian test-takers, self-assessment, TOEFL

Cite this article: Rezvani, R., & Khadivikia, F. (2025). IELTS and TOEFL applicants’ self-assessment and actual test performance: A mixed-methods study of relationship and variation. *Journal of Modern Research in English Language Studies*, *12*(2), 75-98. DOI:10.30479/jmrels.2024.20367.2378

Publisher: Imam Khomeini International University

1. Introduction

In the past twenty years, there has been increasing recognition of the importance of involving students more actively in the learning and assessment process to enhance their future success (Cassidy, 2007; Sambell et al., 2013; Yan, 2020). As a result, there is now greater attention being given to student-oriented assessment methods, such as self-assessment and peer-assessment, despite the challenges and time constraints they may present (Boud & Soler 2016; González-Betancor et al., 2019). It is argued that in order for students to improve their learning, they need to develop the ability to evaluate the quality of their own learning (González-Betancor et al., 2019).

In the field of assessment research, self-assessment (SA) is a practice where individuals assess their own knowledge. This practice emerged in language studies during the 1980s (Oscarson, 2013). SA has received considerable attention and has been presented to learners as an alternative to traditional assessment methods, allowing learners to assess their progress more accurately and understand how to evaluate their future language performance (Çakmak et al., 2023; Zheng et al., 2023). Research has shown several benefits of SA. SA is believed to promote learners' self-regulation as it helps them to set goals and criteria, monitor their performance, reflect on their progress and internalize their learning experience, enabling them to take more responsibility for their own learning (Ajjawi & Boud, 2018; Butler, 2024). SA has also inspired research on whether it can be used as a valid tool for measuring language proficiency (Li & Zhang, 2021). However, there has always been a concern about whether SA instruments truly reflect language performance. Empirical investigations, such as correlational studies between SA scores and scores on external measures like language proficiency tests (e.g., IELTS and TOEFL), have often been conducted to address this concern (Li & Zhang, 2021; Liu, 2021; Ma & Winke, 2019). The association between self-rated scores and actual performance has been found to vary from very weak to very strong. For example, Trofimovich et al. (2016) observed no relationship between speakers' self-ratings of accentedness and comprehensibility and their actual scores.

The review of the literature reveals a generally positive correlation between self-assessment (SA) and language performance measures. However, there is still significant variation in the strength and significance of these correlations across studies. The only study that has explored the correlation between SA and performance on IELTS is the study by Smith (2015). Smith found that participants' self-assessments of their speaking skills accurately predicted their performance on a simulated IELTS speaking task. However, the accuracy of SA decreased among participants with lower proficiency levels. Therefore, further research on this topic is needed.

Although research on SA of language proficiency has expanded and gained more attention from researchers recently, there is still a lack of comprehensive investigations into the validity of SA for assessing performance on high stakes language proficiency tests such as IELTS and TOEFL. The few studies that have explored SA of IELTS and TOEFL, had significant limitations in their design. They were narrow in scope and only assessed performance in specific skills (Richard, 2020; Runnels, 2016; Smith, 2015). Furthermore, these studies (Smith, 2015; Trofimovich et al., 2016) did not provide insights into the reasons behind discrepancies between SA and actual test performance, nor did they investigate test takers' perceptions. Therefore, the primary issue with research on SA and performance in language proficiency tests is emphasis on quantitative approaches and limited assessment of language skills. To address this research gap, the present study aims to explore the following three research questions:

1. Is there a significant relationship between IELTS and TOEFL examinees' self-assessments and their actual performance on the speaking, listening, writing and reading modules?
2. How well do IELTS and TOEFL test takers' self-assessments predict their actual test scores?
3. What factors contribute to the variations between IELTS and TOEFL applicants' self-assessments and their actual performance on the four modules?

2. Literature Review

2.1. Theoretical Background

In the realm of language assessment, SA has become a widely adopted alternative to traditional teacher-led assessment, particularly in formative settings (Fan & Yan, 2017). However, the utility of SA extends beyond formative contexts, as it can also serve as a valuable tool for measuring language proficiency in high-stakes language testing domains (Fan & Yan, 2017). One of the key challenges facing language test developers is the task of gathering and analyzing predictive validity evidence, which sheds light on the degree to which test scores can predict future language performance in these domains (Fan & Yan, 2017). This challenge is compounded by the difficulty in determining the precise elements that contribute to a robust measure of authentic language use (Fan & Yan, 2017). Despite criticisms regarding the reliability and validity of SA, this approach offers a practical and justifiable alternative for assessing language proficiency in testing contexts (Blanche & Merino, 1989). Through the process of constant self-reflection, individuals can develop a nuanced understanding of their own language abilities in authentic settings, often rising above the insights and feedback that external evaluators can provide (Powers & Powers, 2015). Moreover, research has demonstrated the acceptable predictive accuracy of SA across a range of contexts (Blanche & Merino, 1989). While factors such as reliability and validity remain a concern, these can be addressed through the strategic use of tools like detailed contextual information and enhanced content validity of the SA instrument (Suzuki, 2015). By thoughtfully addressing these considerations, language test developers can leverage the unique strengths of self-assessment to effectively measure language proficiency in regular classrooms and high-stakes testing domains.

2.2. Accuracy of Self-Assessment

Research on self-assessment accuracy has yielded diverse findings (Harris & Brown, 2013). While some studies have reported a high level of self-assessment accuracy (Panadero & Romero, 2014), Kun (2016) has demonstrated an overestimation of self-assessment accuracy. Comparing self-assessment results with external criterion measures such as teacher ratings, final grades or objective tests through correlation analyses (Butler & Lee, 2006) has been a common method to measure the accuracy of self-assessment in previous research. Despite inconsistent findings, the main conclusion drawn from these studies underscores the reliability of SA for assessing the proficiency levels of second language learners (Fan, 2016).

2.3. Self-Assessment of Second Language Skills

Researchers have often explored the relationship between self-assessment and language tests, but varying results have been presented. This variability stems from the need to use self-assessment scales carefully and skilfully (Ross, 1998). A meta-analysis conducted by Li and Zhang (2021) revealed that listening skills showed the strongest correlation ($r = 0.486$), followed by reading ($r = 0.451$) and speaking ($r = 0.442$). Writing skills had the weakest correlation ($r = 0.381$), which is consistent with the results reported by Ross (1998). However, other studies have also shown moderate to high correlations, such as 0.75-0.96 in Bachman and Palmer (1989) and 0.61-0.84 in Birjandi and Bolghari (2015).

Speaking assessments have often involved subjective ratings, leading to variations due to different interpretations. Speaking skills have produced correlations ranging from 0.44 to 0.63. For instance, Mahmoodi and Karampour (2019) revealed that there was a significant positive relationship ($r = 0.62$) between meta-cognitive self-regulation and L2 speaking performance of Iranian intermediate EFL learners.

Regarding self-assessment of listening skills, the literature reports varying correlations between learners' self-assessments and other assessments, ranging from no significant correlation to weak or significant correlations. For example, one study (Runnels, 2016) found a weak correlation between TOEIC listening scores and self-assessment ratings. The variation in correlations stemmed from learners' limited experience and subjective assessment criteria.

Results in reading self-assessment have often been varied. Brantmeier (2005) found no correlation between SA and multiple-choice assessments, but observed a correlation with free recall measures of reading comprehension. Similarly, Runnels (2016) discovered a near-negligible, yet negative, correlation between TOEIC scores and CEFR-J reading self-assessment ratings (r = - 0.14). In contrast, Richard (2020) found a moderate correlation between TOEIC scores and mean difficulty ratings of CEFR-J reading (r = 0.47). Azmoode Sis Abad et al. (2024) found no significant difference between diagnostic self- and peer assessment on promoting reading comprehension of Iranian EFL learners (r = .379).

Regarding self-assessment of writing, the literature reveals a positive relationship between SA and writing performance (Alkhowarizmi & Hamdani, 2022; Liu & Brantmeier, 2019; Summers et al., 2019; Wind, 2021; Zheng et al., 2012). Many studies have utilized a cross-sectional research design (Wind & Zólyomi, 2022). However, the strength of these relationships range from weak to moderate. For example, Saito and Fujita (2004) found a weak correlation between SA and teacher assessments (r = 0.07), whereas Weigle (2010) detected moderate positive correlations between SA and teacher assessments (rater 1: r = 0.39, rater 2: r = 0.43). Liu and Brantmeier (2019) observed a significant positive relationship between SA writing and writing production (r = 0.30, p < 0.01). Furthermore, Plonsky and Oswald (2014) revealed a significant positive relationship between SA writing and writing production (r = 0.30, p < .01). Hasnalia et al. (2023) also discovered moderate levels of self-assessment correlations in writing skills.

The literature on the relationship between SA and language proficiency has produced a wide range of empirical findings. Various studies have reported different degrees of correlation between these two constructs, ranging from very low to very high (Blanche & Merino, 1989; Runnels, 2016; Suzuki, 2015). This lack of consistency in the observed relationships highlights the complex and multifaceted nature of the connection between self-perceived language abilities and objective performance on proficiency tests. To address these inconsistencies and gain a more comprehensive understanding of the SA-proficiency relationship, researchers have emphasized the importance of using standardized and widely recognized assessment instruments (Runnels, 2016). Based on the recommendations from existing literature, this study intends to further explore the relationship between self-perceived language abilities and objectively measured language proficiency.

3. Method

3.1. Participants

A mock test was given to randomly selected IELTS and TOEFL applicants who were enrolled in preparation courses to ensure that they had a similar level of overall language proficiency. Ultimately, 81 participants were recruited from IELTS and TOEFL preparation centers using convenience and snowball sampling methods. The data for this study were collected from a sample of 81 applicants who had taken either the IELTS or TOEFL tests. Among these participants, 51 were IELTS applicants and 30 were TOEFL applicants. The age of the participants ranged from 18 to 35. In terms of gender, there were 50 male applicants and 31 female applicants. Among the male participants, 32 were IELTS test takers and 18 were TOEFL test takers. On the other hand, among the 31 female participants, 19 were IELTS applicants and 12 were TOEFL test takers. Regarding educational background, 6 participants were Ph.D. candidates, 25 had an MA degree, 40 had a BA degree, and 10 were college students. All of the participants had prior experience with both exams, having taken them at least once. The applicants were selected to have a similar level of proficiency at the intermediate and advanced levels.

3.2. Instruments

Two instruments were employed in the current study; one was an SA questionnaire; the other was semi-structured interviews. In the questionnaire, the first two items were related to demographics (gender and age). The first three questions tapped the participants' background information (the way they had learned English, their contacts with native speakers and length of those contacts). The next three questions were designed to collect data on their assessment history (previous experience in IELTS, TOEFL and mock tests, participation in preparation courses and frequency of it and SAs, if any). The answers to these questions were employed as a tool for homogenization of the participants. Finally, they were asked to write their SA and AT scores on a table. We used the recorded scores on the table as the source of correlation and regression analyses (see appendix A).

The interviews had two parts (see appendix B). In the first part, the questions elicited information on the participants' (1) language skills and SAs, (2) and details of when, how and times of their SAs. In the second part, they gave detailed answers concerning (1) the perceived general differences and similarities between the results of their ATs and SAs, (2) the most expected and unexpected AT and SA results and (3) the reasons behind the discrepancies between their SA and AT results. The achieved data were used as the source of qualitative data after classifying and coding them. The questionnaire and interview questions were piloted with a small sample of similar test takers to identify and address any potential difficulties that they might face in understanding them. The sample consisted of 2 male and female IELTS applicants. After they received the questionnaire and interview questions, the purpose of the study was explained to them. They completed the questionnaire prior to the interview. They were asked to highlight any ambiguities or flaws they had identified in the questions. Based on their feedback, the questionnaire and interview questions were revised.

3.3. Procedure

The questionnaire was made accessible to the participants through e-mail. A consent form was also sent along with it to be studied and signed by them if they consented to participate in the study. To avoid any ambiguities, the participants were briefed by the email about the questionnaire items and also the purpose of the study.

After receiving the filled-out questionnaires, we randomly invited 40 applicants by a second email to participate in face-to-face or telephone interviews. Thirty of them agreed to take part in the interviews. These participants were given the option to answer questions either in Persian or English to facilitate the expression of their experiences. All the interviews were conducted individually by the second author taking 15 to 25 minutes. All the interviews were recorded, transcribed and then thematically analyzed to answer the research questions.

3.4. Data Analysis

In order to analyze the quantitative data, the scores obtained from the respondents' self-assessments and actual test scores were compiled into a dataset and entered into SPSS for correlation and regression analyses. This process aimed to provide a numerical representation of the relationship between the participants' SA and actual test scores. For the qualitative data analysis, an inductive thematic content analysis approach was utilized which involved a three-step process consisting of data familiarization, code generation and theme extraction. The analysis of the transcripts entailed a meticulous and iterative examination of the collected data to identify patterns and similarities. To ensure the trustworthiness of the findings, memos and field notes were also incorporated in the thematic analysis process (Riazi et al., 2023). Moreover, to maintain consistency in data coding and analysis, the interview data were initially coded and analyzed by the second researcher, with 20% of the data being reanalyzed by the first researcher. The initial inter-coder reliability stood at 85%. Any discrepancies were addressed through discussion and resolution, followed by a reanalysis of a new portion (10%) of the data by both researchers, resulting in an agreement rate of 96%. Subsequently, the data were reanalyzed by the second researcher for any necessary amendments to the findings (see Rezvani & Miri, 2021).

4. Results and Discussion

4.1. Results

4.1.1. Quantitative Results

Firstly, the normality of the data was assessed using the Kolmogorov-Smirnov Test. As the data followed a normal distribution, statistical methods were applied for analysis. The mean values of the IELTS SA and AT scores ranged from 5.94 to 7.00, respectively, indicating a difference between the SA and AT scores. The SA and AT scores for reading had the smallest difference, suggesting that test-takers encountered the least difficulty in these sections. Conversely, the SA and AT scores for listening had the largest difference, indicating that test-takers faced the highest difficulty in this module. Additionally, the mean values of the TOEFL SA and AT scores ranged from 23 to 26, respectively. For more details, refer to Table 1.

Table 1
The Descriptive Statistics of SA and AT Scores of IELTS and TOEFL Tests

Test Module	Assessment	Mean	SD	P
Speaking	I SA	6.17	0.74	0.01
	I AT	6.11	0.68	
	T SA	25.00	1.22	
	T AT	26.00	1.77	
Listening	I SA	5.94	0.58	0.01
	I AT	5.81	0.74	
	T SA	26.33	0.99	
	T AT	23.81	1.02	
Writing	I SA	6.38	0.69	0.01
	I AT	6.38	0.71	
	T SA	26.03	1.21	
	T AT	23.66	0.75	
Reading	I SA	6.08	0.75	0.01
	I AT	7.00	0.44	
	T SA	24.80	0.92	
	T AT	25.03	1.06	

The first research question addressed by this study was whether there was a significant correlation between the self-assessments of IELTS and

TOEFL examinees and their actual performance on the speaking, listening, writing and reading modules. The results of Pearson product-moment correlations revealed moderately high positive correlations between the test takers' self-assessments and their actual test scores. Moderate correlations were found between IELTS and TOEFL speaking and reading, as well as between IELTS writing and TOEFL listening and writing. The highest correlation was observed for IELTS reading (r = 0.780; p < 0.01, two-tailed), while the lowest correlation was found for TOEFL listening (r = 0.572; p < 0.01, two-tailed). For more details, please refer to Table 2.

Overall, the correlation values suggest that IELTS test-takers performed better in both their self-assessments and their actual test performance compared to TOEFL test-takers. However, it is important to note that this difference was not particularly significant. Another important consideration is that although the Pearson values indicated high positive correlations, it does not necessarily imply a causal relationship between the two variables, namely self-assessed and actual test scores.

Table 2

Correlations among SA and AT in IELTS and TOEFL Speaking, Listening, Writing and Reading Modules

Test Module	Assessment	Pearson Correlation	Sig. 2-tailed
Speaking	IELTS	0.762	0.01
	TOEFL	0.738	
Listening	IELTS	0.605	0.01
	TOEFL	0.572	
Writing	IELTS	0.728	0.01
	TOEFL	0.648	
Reading	IELTS	0.780	0.01
	TOEFL	0.741	

4.1.2. Applicants' Prediction Accuracy

The second research question focused on whether IELTS and TOEFL applicants accurately predicted their performance on the test. In order to explore this question, a linear regression analysis was conducted to gain a more comprehensive understanding of the observed correlation patterns.

The findings of the regression analysis revealed that, with the exception of the listening module and TOEFL writing, the self-assessments significantly accounted for the variability in the actual test performance. Specifically, the IELTS reading module had the highest contribution to the variability in the actual test score, explaining 61.2 percent of the variation. This suggests a significant linear association between self-assessments and actual test performance, as indicated by an R^2 value of 0.612 (adjusted R^2 = 0.596). The second highest contributor to the variability was the TOEFL reading module, with an R^2 value of 0.583 (adjusted R^2 = 0.522). Following that, the IELTS speaking module had an R^2 value of 0.581 (adjusted R^2 = 0.564). The TOEFL speaking module had a relatively similar R^2 value to the IELTS speaking module, with an R^2 value of 0.560 (adjusted R^2 = 0.527). Subsequently, the IELTS and TOEFL writing modules had R^2 values of 0.549 (adjusted R^2 = 0.531) and 0.430 (adjusted R^2 = 0.388), respectively. Lastly, the listening module had the lowest predictive value among the four modules, with an R^2 value of 0.387 for IELTS (adjusted R^2 = 0.361) and 0.339 for TOEFL (adjusted R^2 = 0.290) (Refer to Table 3 for further details).

Table 3

Regression Models for Predicting Applicants' Accuracy in IELTS and TOEFL Speaking, Listening, Reading, and Writing Modules

Test module	Assessment	Adj. R^2	R^2	Standard Error of Estimate
Speaking	IELTS	0.564	0.581	0.451
	TOEFL	0.527	0.560	1.224
Listening	IELTS	0.361	0.387	0.592
	TOEFL	0.290	0.339	0.860
Writing	IELTS	0.531	0.549	0.487
	TOEFL	0.388	0.430	0.593
Reading	IELTS	0.596	0.612	0.284
	TOEFL	0.522	0.583	0.714

Note. $p < 0.01$.

4.1.3. Qualitative Results

The applicants' responses to interview questions are presented and discussed below. Common themes and subthemes are presented in Table 4.

Table 4
IELTS and TOEFL Applicants' Perceived Sources of Variation

Themes	Subthemes
Experience	Familiarity with IELTS/TOEFL test taking strategies Familiarity with test structure/constituents The effect of test situation
Linguistic factors	Linguistic abilities Personal attributes
Feedback	Attitudes The role of preparation courses
Background knowledge	Familiarity/Unfamiliarity with subject matter
Psychological factors	Anxiety

The emerged themes revealed that some non-linguistic factors affected the candidates' real performance contrary to their predictions. The perceived factors either contributed to successful SA or inhibited it. The findings also support previous research which though English proficiency has a major role to play in satisfactory performance in language tests, factors other than language proficiency may have contributed to success or failure in those tests. Some candidates stated that despite the fact that they had predicted that they would obtain high actual test scores based on their self-assessments, which demonstrated their ability to speak or write in English, they did not manage to acquire their desired scores, relating it to lack of test experience. The effect of test experience was evident in their responses when they referred to items like applying test-taking strategies, coping with test structure and or constituents, and the effect of test situation.

> When you take IELTS, … having good scores depends very much on strategies even if you have high English abilities.

However, for a number of participants, familiarity with test-taking strategies was the most efficient factor in their test-taking success. Also, they made references to teachers' guides and advice on how to use testing materials:

> For improving our speaking, the teacher designed questions and gave us texts to read and requested to discuss questions related to it.

Previous experience in the test was referred to by the frequent use of words like "structure of the test" and "awareness of test tasks":

> I have taken IELTS two times, therefore, I held an entirely realistic and comprehensive vision of the IELTS components.

Poor listening scores were linked to the “perplexing” structure of the test items mirrored by recurring use of items such as “difficulty in understanding native accent”, “organising information”, “confusing”, “short period of time” and “academic questions”.

Also mentioned in the comments was the unexpected difference between the actual test situation and simulated test-taking experiences (“The context of the actual exam is very different from mock tests or exam practices”.).

While “test-taking experiences” was the keyword in the participants’ responses, in some cases, similarities between test items in the test and the education system (“One of the questions requires to recognize the items in the responses that are not true according to the passage”.) and the positive effect of academic knowledge (“I had a good working knowledge of language skills … because of … writing academic papers”.) were also observed.

The effect of anxiety was especially noticeable in the test-takers’ writing (“I was worried about lack of time to complete the writing task”.), speaking (“I felt my mind was empty and looked for suitable structures … because of the environment of the exam”.), and listening (“I had feelings of anxiety because the speakers in the conversations spoke very fast”.).

For some respondents, the difference between real-life communications and the speaking and listening items of the test were cited, as in real-life, the speaker or listener, “receives or gives feedback” (“… you receive no comment or reaction on your words”.) and is not in a rush to throw in a response, while in the test tasks, test-takers may “feel anxious” about the “time limit”.

Linguistic abilities and personal attributes were among the factors expressed ubiquitously by the test-takers, and were reflected by phrases like lack of fluency (“… I have not listened enough to native accent speakers because instructors in Iran are all non-natives”.), and practicing unplanned conversation and designing exercises.

The writing module appeared to play a negative role in the mismatch between before-exam assessments and actual exam performance:

> Writing needs very formal words and structures and it was difficult to find good items.

Personal attributes or test strategies contributed positively (“I tried to speak a lot before the test about topics which were likely to be asked”.) or negatively (“I could not identify key words in the voices especially when the speakers referred to a particular time or a shift from one topic or time to another”.).

The difficulty of the items being higher than expected were also noticed in the responses as, for instance, one participant stated that it was her first time that she took the test and she was not aware of the difficulty of the actual test enough.

> For lack of coherence, I repeated sentences for several times … and for writing, I could not develop the topic sufficiently.

For some of the participants, it was their attitudes that affected their test performance and the way in which they had responded to test items.

> I think I had wrong impressions about the things that were emphasized in the actual exam.

Some of the participants spoke specifically of the effect of their "preparation procedures" on their ability to self-assess ("I expected too much from a two-month preparation course which was too short to meet expectations".). Some participants argued that although preparation courses were very important in preparing for the exam, relying solely on "preparation courses and mock tests" would not yield intended results. Other examinees, however, spoke positively of the effect of preparation courses on their test-taking skills.

> The instructor's emphasis on appropriate behaviour reduced the difference between my SA and IELTS score.

Uncertainties during the test, holding positive beliefs about the nature of the test, especially as compared to IELTS, were also tipped as potential factors affecting test performance. Some of the respondents argued that when they asked their experienced peers, they strongly advised "memorisation of fixed expressions and chunks" as an important strategy in TOEFL speaking and writing. However, the participants expressed worries over being unsure of how to respond to test items due to uncertainties about the scoring policies of the TOEFL test:

> I was good at writing long texts but I was doubtful because I didn't know whether examiners accept a long or short one.

Interestingly, some of the applicants believed that it was easier to predict performance in the IELTS test because the TOEFL test is inclined towards discipline specific items and favour test-takers familiar with academic topics.

> It is pretty easier to train yourself for TOEFL because IELTS requires higher levels of cognitive abilities.

Background characteristics like familiarity with subject matter did not appear to be an as much effective factor as experience, linguistic factors or feedback on IELTS. The general argument was that relying solely on background characteristics, particularly subject matter, would not yield accurate test results.

> I relied on my academic English experience but I failed to get the band scores needed.

In contrast, for some TOEFL participants, accurate test results were at least partially related to the degree of familiarity with a subject matter. It means that the TOEFL test was fairly discipline specific compared to the IELTS test:

> I have taken both IELTS and TOEFL. … my performance in the TOEFL test did not represent my abilities because the topics were too much concerned with a particular discipline.

4.2. Discussion

In recent decades, there has been an increasing emphasis on self-assessment from a variety of perspectives. These studies are often conducted to gauge progress, typically using can-do statements. As a result, the aim of this study was to quantitatively examine the correlation between self-assessed and actual test performance among applicants taking the IELTS and TOEFL exams. Additionally, this study sought to qualitatively explore the factors that hinder accurate self-assessment of test takers' performance in these two high-stakes tests, particularly with regard to Iranian applicants. In order to provide a more comprehensive and direct discussion of the findings on self-assessment and actual test performance of IELTS and TOEFL applicants, the researchers sought the respondents' perspectives on their self-assessed and actual test scores, as well as the sources of variation between these two measures.

In relation to the first research question, the findings in the literature are varied compared to this study. Specifically, high correlations were found between IELTS and TOEFL speaking and reading, as well as IELTS writing. Additionally, moderate correlations were observed between IELTS and TOEFL listening, and TOEFL writing. The highest correlation was found in IELTS reading, while the lowest one was in TOEFL listening. However, it is essential to emphasize that despite the presence of high correlations, it should not be assumed that there is a direct causal link between self-assessed and actual test scores.

In a meta-analysis by Li and Zhang (2021), in which they explored the correlation between SA and language performance in 67 papers, an overall correlation of 0.466 was revealed compared to the overall relationship of z = 0.472 indicated in another meta-analysis by León et al. (2023). In this study, moderate to high relationships were observed between self-assessed and actual test scores of IELTS and TOEFL applicants across the four modules. The speaking correlation discovered by the meta-analysis was r = 0.442 whereas the current study revealed a correlation of r = 0.762 and r = 0.738 for IELTS and TOEFL, respectively.

In another study (Mahmoodi & Karampour 2019), a significant positive correlation (r = 0.62) was obtained between meta-cognitive self-regulation and L2 speaking performance of Iranian intermediate EFL learners. Summers et al. (2019) also obtained a moderate relationship (r = 0.44) between speaking SA against ACTFL Can-Do Statements.

In terms of the listening skill, this study's relationships were r = 0.605 (IELTS) and r = 0.572 (TOEFL), which were the lowest ones in line with the weak correlation (r = 0.32) between TOEIC listening scores and self-assessment ratings produced in Runnels's study (2016) and in contrast to Li and Zhang's (2021) correlation of r = 0.486, which was the strongest one among other skills.

In terms of writing skills, the relationships found in the literature vary, with correlations ranging from r = 0.30 (Liu & Brantmeier, 2019), r = 0.381 (Li & Zhang, 2021), r = 0.47 (Summers et al., 2019), and r = 0.647 (Mohammadi et al., 2024). However, the correlations discovered in our study exceeded those reported in the literature, with IELTS scoring r = 0.728 and TOEFL r = 0.648.

Moreover, the correlations for reading skills, which are the highest achieved in our research, are r = 0.780 and r = 0.741 for IELTS and TOEFL, respectively. In contrast, the literature presents correlations of r = - 0.14 (Runnels, 2016), r = 0.451 (Li & Zhang, 2021) and r = 0.47 (Richard, 2020).

Regarding the second research question, the regression analyses revealed that the self-assessments played a significant role in explaining the variability in the actual test performance, with the exception of the listening module and TOEFL Writing. Among the different modules, IELTS Reading had the greatest impact on the test scores. This suggests that there was a strong linear relationship between self-assessments and actual test performance. The second most influential module was TOEFL Reading, followed by IELTS Speaking, while TOEFL Speaking had a slightly lower R^2 value. IELTS and TOEFL Writing were the next modules. Finally, the listening module had the least predictive power among the four modules.

The findings in this study broadly align with previous research on self-assessment in the context of language tests. Skehan (2014) argues that personal traits such as motivation, anxiety, ambiguity tolerance, and others can also impact test takers' performance. These construct-irrelevant factors are potential sources of test bias that can distort the obtained scores, making them unrepresentative of the underlying ability that a language test aims to measure and compromising the integrity of the testing process (Takala & Kaftandjieva, 2000). Cotton and Conrow (1998) assert that elevated levels of English proficiency, as measured by the IELTS test, do not inevitably translate into academic accomplishment.

The literature supports this study's finding that knowledge in specific content areas may contribute to variations in language test performance. These background characteristics encompass various factors, including cultural background, familiarity with specific content areas, cognitive style, native language, cognitive ability, gender, and age (Kunnan, 2007).

The literature presents mixed results regarding the relationship between IELTS test performance and background knowledge. Each of the four language skills assessed by IELTS, reading, listening, speaking, and writing, is strongly correlated, yet distinct. Consequently, the impact of test takers' background characteristics may differ across these skills (Manna & Yoo, 2015).

However, as a crucial aspect of test preparation courses, awareness of assessment criteria seems to have been overlooked by preparation courses in Iran. It is important to note, though, that there is a lack of empirical evidence (Green, 2013) to support the idea that focusing extensively on these practices would yield desired outcomes. Alderson and Hamp-Lyons (1996) argue, based on their observation of TOEFL preparation classes, that if teachers carefully select appropriate content and methods for test preparation, their TOEFL teaching could result in positive washback.

Test anxiety during the actual test reduces test-takers' attention and increases the likelihood of errors (Cassady & Johnson, 2002; Ohata, 2005). However, Chapelle et al. (2011) suggest that, for some students, some levels of anxiety can be beneficial. It can motivate longer study periods and promote careful attention to exam questions. Stricker et al. (2004) identified that students generally held positive attitudes towards computer-based testing.

5. Conclusion and Implications

In conclusion, this study sheds light on the intricate nature of self-assessment accuracy in language proficiency tests such as the IELTS and TOEFL. The findings underscore the complexity of the relationship between self-assessment and actual test performance, particularly in the speaking, reading, and writing modules. While there is a moderate to high correlation between self-assessment and test scores in these modules, the variability in predictive power, especially in the listening section, emphasizes the need for a more nuanced understanding of self-assessment processes. Furthermore, the study highlights the influence of non-linguistic factors on self-assessment accuracy and test performance. Test experience, for instance, plays a crucial role in shaping individuals' perceptions of their language abilities and their preparedness for the test. With regard to psychological factors such as test anxiety, motivation and self-efficacy test-takers who experience high levels of anxiety may underestimate their abilities or perform below their actual potential on the test, leading to discrepancies between their self-assessment and test scores. Conversely, individuals with strong motivation and self-efficacy beliefs may provide more accurate self-assessments and achieve better test outcomes. Concerning the effect of preparation courses and study strategies, engaging in targeted language development activities, familiarizing oneself with test formats and receiving guidance from experienced instructors can enhance test-takers' understanding of their strengths and weaknesses.

Overall, by increasing test-takers' awareness of these non-linguistic factors and emphasizing comprehensive language development, there is potential to enhance self-assessment accuracy and ultimately improve language test outcomes.

It is important to acknowledge the limitations of the current study. The number of participants was relatively small, and the study was limited to applicants in Iran. Including more applicants from both tests or even from diverse contexts would have provided a broader perspective on the variations between IELTS and TOEFL test takers' self-assessment and actual test performance. Future research should investigate the self-assessment of test-takers with a larger number of participants from different contexts and with test-takers of other language proficiency tests. Additionally, future studies could explore the perceptions of instructors, examiners, and test designers to provide a more holistic view of the factors influencing self-assessment and actual test performance. Future research could delve deeper into how cultural backgrounds and language learning experiences may influence test-takers' self-assessment tendencies and test performance. Finally, by exploring the cultural dimensions of self-assessment, researchers can enhance the generalizability and applicability of their findings.

Acknowledgements

We thank the IELTS and TOEFL training institutes for helping us with data collection and applicants who collaborated with us at different stages of the work.

References

Ajjawi, R., & Boud, D. (2018). Examining the nature and effects of feedback dialogue. *Assessment & Evaluation in Higher Education*, *43*(7), 1106-1119. https://doi.org/10.1080/02602938.2018.1434128

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing*, *13*(3), 280-297. https://doi.org/10.1177/026553229601300304

Alkhowarizmi, A., & Hamdani, H. (2022). The effect of using self-assessment technique towards EFL students' writing skill. *Edulitics (Education, Literature, and Linguistics) Journal*, *7*(2), 88-100.

Azmoode Sis Abad, M., Kiany, Gh. R., & Abassian, Gh. R. (2024). On the effect of diagnostic self- and peer-assessment on reading comprehension: Examining EFL learners' diagnostic rating accuracy across various genres. *Journal of Modern Research in English Language Studies,* *11*(2), 177-202. https://doi.org/10.30479/jmrels.2023.18703.2204

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*(1), 14-29.

Birjandi, P., & Bolghari, M. S. (2015). The relationship between the accuracy of self- and peer-assessment of Iranian intermediate EFL learners and their learning styles. *Theory and Practice in Language Studies*, *5*(5), 996-1006.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning, 39*(3), 313-338.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400-413.

Brantmeier, C. (2005). Nonlinguistic variables in advanced second language reading: Learners' self-assessment and enjoyment. *Foreign Language Annals*, *38*(4), 494-504.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal*, *90*(4), 506-518. https://doi.org/10.1111/j.1540-4781.2006.00463.x

Butler, Y. G. (2024). Self-assessment in second language learning. *Language Teaching*, *57*(1), 42-56.

Çakmak, F., Ismail, S. M., & Karami, S. (2023). Advancing learning-oriented assessment (LOA): Mapping the role of self-assessment, academic resilience, academic motivation in students' test-taking skills, and test anxiety management in Telegram-assisted language learning. *Language Testing in Asia*, *13*(1), 1-19.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270-295. https://doi.org/10.1006/ceps.2001.1094

Cassidy, S. (2007). Assessing inexperienced students' ability to self-assess: Exploring links with learning style and academic personal control. *Assessment & Evaluation in Higher Education*, *32*(3), 313-330.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports*, *1*(4), 72–115.

Fan, J., & Yan, X. (2017). From test performance to language use: Using self-assessment to validate a high-stakes English proficiency test. *The Asia-Pacific Education Researcher*, *26*, 61-73.

Fan, J. J. (2016). The construct and predictive validity of a self-assessment scale. *Papers in Language Testing and Assessment*, *5*(2), 69-100.

González-Betancor, S. M., Bolívar-Cruz, A., & Verano-Tacoronte, D. (2019). Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education*, *20*(2), 101-114.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39-51.

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education*, *36*, 101-111. https://doi.org/10.1016/j.tate.2013.07.008

Hasnalia, R., & Rifli, N. R. (2023). Self-assessment of English writing skills. *Pedagogical Research Journal*, *1*(1), 7-12.

Kunnan, A. J. (2007). Test fairness, test bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109-112. https://doi.org/10.1080/15434300701375865

Kun, A. I. (2016). A comparison of self- versus tutor assessment among Hungarian undergraduate business students. *Assessment & Evaluation in Higher Education*, *41*(3), 350-367. https://doi.org/10.1080/02602938.2015.1011602

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review*, *35*(4), 106.

Li, M., Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189-218. https://doi.org/10.1177/0265532220932481

Liu, J. (2021). Correlating self-efficacy with self-assessment in an undergraduate interpreting classroom: How accurate can students be? *Porta Linguarum: Revista Internacional De Didáctica De Las Lenguas Extranjeras*, (36), 9-25.

Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System*, *80*, 60-72.

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals, 52,* 66-86. https://doi.org/10.1111/flan.12379

Mahmoodi, M. H., & Karampour, F. (2019). Relationship between Iranian intermediate EFL learners' foreign language causal attributions, meta-cognitive self-regulation and their L2 speaking performance. *Journal of Modern Research in English Language Studies*, *6*(2), 77-53. https://doi.org/10.30479/jmrels.2019.11253.1406

Manna, V. F., & Yoo, H. (2015). Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English as a second language (ESL) test population: A factor analytic approach. *ETS Research Report Series, 2015*(2), 1-23. https://doi.org/10.1002/ets2.12072

Mohammadi, R., Ghanbari, N., & Abbasi, A. (2024). Perceptual (mis) matches between learners' and teachers' rating criteria in the Iranian EFL writing self-assessment context. *International Journal of Language Testing*, *14*(1), 150-165.

Ohata, K. (2005). Potential sources of anxiety for Japanese learners of English: Preliminary case interviews with five Japanese college students in the US. *TESL-EJ, 9*(3), 1-21.

Oscarson, M. (2013). Self-assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development* (pp. 712-729). John Wiley & Sons, Inc.

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 133-148.

Plonsky, L., & Oswald, F.L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC listening, reading, speaking, and writing tests to predicting performance on real-life English language tasks. *Language Testing*, *32*(2), 151-167.

Richard, J. P. J. (2020). Investigating CEFR-J Self-assessment and TOEIC listening and reading scores. *The Global Management*, *3*, 21-32.

Rezvani, R., & Miri, P. (2021). The impact of gender, nativeness, and subject matter on the English as a second language university students' perception of instructor credibility and engagement: A qualitative study. *Frontiers in Psychology*, *12*, 1-14. https://doi.org/10.3389/fpsyg.2021.702250

Riazi, A. M., Rezvani, R., & Ghanbar, H. (2023). Trustworthiness in L2 writing research: A review and analysis of qualitative articles in the Journal of Second Language Writing. *Research Methods in Applied Linguistics*, *2*(3), 1-14.

Riazi, A. M., Ghanbar, H., & Rezvani, R. (2023). Qualitative data coding and analysis: A systematic review of the papers published in the Journal of Second Language Writing. *Iranian Journal of Language Teaching Research*, *11*(1), 25-47.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1-20.

Runnels, J. (2016). Self-assessment accuracy: Correlations between Japanese English learners' self-assessment on the CEFR-Japan's can do statements and scores on the TOEIC. *Taiwan Journal of TESOL*, *13*(1), 105-137.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, *8*(1), 31-54.

Sambell, K., McDowell, L., & Montgomery, C. (2013). Developing students as self-assessors and effective lifelong learning. In K. Sambell, L. McDowell, & C. Montgomery (Eds.), *Assessment for learning in higher education* (pp. 120-146). Routledge.

Skehan, P. (2014). Individual differences in second language learning. In E. Kerz, & D. Wiechmann (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 394-406). Routledge.

Smith, S. (2015). Accuracy in speaking self-assessment among Japanese-speaking English learners and its implications. *Polyglossia: The Asia-Pacific's Voice in Language and Language Teaching*, *27*, 41-55.

Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based test of English as a foreign language. *Computers in Human Behavior*, *20*(1), 37-54. https://doi.org/10.1016/S0747-5632(03)00046-3

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an intensive English program. *System*, *80*, 269-287.

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, *32*(1), 63-81.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323-340. https://doi.org/10.1177/026553220001700303

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122-140.

Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335-353.

Wind, A. M. (2021). Nonlinearity and inter- and intra-individual variability in the extent of engagement in self-reflection and its role in second language writing: A multiple-case study. *System*, *103*, 102672.

Wind, A. M., & Zólyomi, A. (2022). The longitudinal development of self-assessment and academic writing: An advanced writing program. *Language Learning in Higher Education*, *12*(1), 185-207.

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education*, *45*(2), 224-239.

Zheng, H., Huang, J., & Chen, Y. (2012). Effects of self-assessment training on Chinese students' performance on college English writing tests. *Polyglossia*, *23*, 33-42.

Zheng, C., Wang, L., & Chai, C. S. (2023). Self-assessment first or peer-assessment first: Effects of video-based formative practice on learners' English public speaking anxiety and performance. *Computer Assisted Language Learning*, *36*(4), 806-839.

Appendices

Appendix A: The Questionnaire Items

Background questions (The respondents' information will be kept confidential and used only for doing research.)

Age: Gender:

1. How did you learn English, self-study, book, institute, private tutors, etc.?
2. Did you ever have a contact with native speakers or travel or stay abroad?
3. If yes, how many times and how long did that contact or travel occur and take?
4. How many times did you take the actual test?
5. Did you attend any preparation courses? If yes, how many times and how long did you attend? What type of preparation course was it?
6. Did you take any mock or tests? If yes, how? Did you do that on your own or it was done by institutes?

7. Please insert the self-assessed and actual test scores in the following table.

Skills	SA scores	AT scores
Speaking		
Listening		
Writing		
Reading		

Appendix B: Interview Questions

1. Did you ever assess your skills, speaking, listening, reading, and writing?
2. If yes, which skills? And:
    a. When?
    b. How?
    c. How many times did you carry out the assessment(s)?
    d. What were the results?

3. Based on what you obtained on the actual exam(s), how similar or different were your self-assessments and the actual test results?
4. Which one was different from your own assessments/expectations? Which one was most unexpected?
5. Concerning speaking, listening, reading, and writing, why do you think the actual test results were close to, the same as, different, or so different from your self-assessments?