Tabaran Institute of Higher Education
*International Journal of Language Testing*
Vol. 12, No. 2, October 2022

ISSN 2476-5880

# Why IELTS Candidates Score Low in Writing: Investigating the Effects of Test Design and Scoring Criteria on Test-Takers' Grades in IELTS and World Englishes Essay Writing Tests

Sajjad Arefsadr[1]*, Esmat Babaii [2], Mohammad Reza Hashemi [3]

Received: March 2022

Accepted: June 2022

**Abstract**

This study explored possible reasons why IELTS candidates usually score low in writing by investigating the effects of two different test designs and scoring criteria on Iranian IELTS candidates' obtained grades in IELTS and World Englishes (WEs) essay writing tests. To this end, first, a WEs essay writing test was preliminarily designed. Then, 17 Iranian IELTS candidates wrote two essays on the same topic, one under the IELTS test condition and one under the WEs test condition. Each of the 34 obtained essays was scored six times, three times based on IELTS scoring criteria, each time by a different rater, and then, three times based on WEs scoring criteria. The results of repeated-measures ANOVA showed that test design and scoring criteria had significant effects on essay grades. The study concludes that some of the reasons why IELTS candidates usually score low in writing may be rooted in the test design and scoring criteria of the IELTS essay writing test, not necessarily in IELTS candidates' weaknesses in writing. The implications of the study focus on the importance and relevance of the results to IELTS candidates, international students, and the future of assessing writing in World Englishes contexts.

*Keywords*: authenticity; communicative proficiency; scoring criteria; test design; WETOAW

## 1. Introduction

One of the most famous large-scale high-stakes English proficiency tests is the International English Language Testing System (IELTS), on which potential candidates usually spend a great deal of time, energy, and money. Understandably, scoring well in IELTS is important to IELTS candidates (Pearson, 2019), especially for the ones who need an IELTS certificate for immigration or study-abroad purposes (Green, 2007).

However, the IELTS academic writing test seems to be one of the major barriers to getting the overall band score that most IELTS candidates need (usually at least 6 or 6.5, see Read, 2015). As reported on the IELTS official website, IELTS candidates usually obtain lower scores in writing than in the other three language skills (IELTS, 2021a). For

[1] Ph.D. Candidate in Applied Linguistics, Department of Foreign Languages, Kharazmi University, No. 49, Mofateh Street, Tehran, Iran. Tel: +98-21-88306651, Email: Std_Sajjad.Arefsadr@khu.ac.ir

[2] Professor in Applied Linguistics, Department of Foreign Languages, Kharazmi University, No. 49, Mofateh Street, Tehran, Iran. Tel: +98-21-88306651, Email: babai@khu.ac.ir

[3]Assistant Professor in Applied Linguistics, Kharazmi University, No. 49, Mofateh Street, Tehran, Iran. Tel: +98-21-88306651, Email: Hashemi_ili@yahoo.com

example, as the IELTS official website (www.ielts.org) reports, in 2019, the academic mean performance of IELTS candidates was lowest in academic writing (compared to speaking, reading, or listening), with males scoring 5.6 and females scoring 5.7 on average (IELTS, 2021a). And, Iranian candidates seem to face the same problem as they scored lowest in writing (5.8, on average) in 2019, compared to reading, listening, and speaking (IELTS, 2021a). The problem with writing, then, seems prevalent among IELTS candidates, making it necessary to address it.

Before addressing this problem, it is to be noted that the reason why we have focused on the essay writing test of IELTS in this study is that, in scoring, “Task 2 contributes twice as much as Task 1 to the writing score” (IELTS, 2018, p. 38) and, therefore, has more contribution to the final band score for writing. Moreover, because of the same scoring criteria and identical format, both IELTS academic and general training essay writing tests have been considered the same in this study, and essays written in these tests are referred to as *IELTS essays* for ease of discussion. We also assume that the reader is familiar with the IELTS academic writing test.

Now, to see why IELTS candidates usually score low in writing, one can see what problems IELTS candidates may have with the IELTS writing tests. However, to put it more in favor of IELTS candidates, one can also see what problems IELTS writing tests may create for test-takers. In other words, the problem of scoring low in writing may not be entirely IELTS candidates’ fault. The IELTS writing tests themselves may have some features that make it difficult for candidates to score well in them, and these features may be rooted in the design and scoring of the IELTS essay writing test.

To see if the *test design* and *scoring criteria* of the IELTS essay writing test contribute to IELTS candidates’ problem of scoring low in writing, they can be compared and contrasted with those of a newly designed essay writing test. This new test that we have preliminarily designed for this study is mainly based on what matters most in *World Englishes* (WEs) contexts, that is, communication (Hu, 2021), and we have called the essays written in this new test *WEs essays*. To shed some light on the reasons why IELTS candidates usually score low in writing and if their essays can obtain better scores if written under different test conditions and scored by different scoring criteria (specified in this study), the following research question has been formed.

Is there a statistically significant difference among Iranian IELTS candidates’ obtained grades in essay writing tests when the essays are written under IELTS and World Englishes test designs and scored by IELTS and World Englishes scoring criteria?

## 2. Review of Literature

One of the main factors that can determine the test design and scoring criteria of a test is the construct of a test (Bachman, 1990; Bachman & Palmer, 1996). Essentially, one should start from defining the construct of a test to specifying its test design and scoring criteria in a way that best serves the construct of the test as much as practicality issues allow it (for aspects of test usefulness see Bachman & Palmer, 1996). Thus, to understand the test design and scoring criteria of the IELTS essay writing test, a clear statement of the construct of this test can help to understand what is assessed by the test and how it is scored. However, as Weigle (2002) noted, official documentation that has been published by IELTS does not give a clear definition of the IELTS academic writing construct, in general, and its essay writing test, in particular. Although Weigle noted this in 2002, since then it does not seem to have been published any official account of what *exactly* the academic writing construct of IELTS is. Only what test-takers are supposed to do in the IELTS academic writing test has been explained in, for example, IELTS handbooks, IELTS preparation books, and on the IELTS official website.

To have a better picture of the construct of the IELTS essay writing test, the descriptions of the test should be taken into consideration. To save space, we refer the reader to the IELTS official website (www.ielts.org) for descriptions of task type and format, task focus, and scoring criteria (see IELTS, 2021b). The descriptions provided on the IELTS official website show that it is important for test-takers to write a discursive essay in “an academic or semi-formal/neutral style”, provide “a full and relevant response”, allocate no more than 40 minutes to the essay, write at least 250 words, avoid a very long essay, write relevantly and on topic in “full, connected text”, avoid plagiarism, and avoid copying “directly from the question paper” (IELTS, 2021b). All this can have some reflections on the way the construct of the test is defined.

Moreover, as Weigle (2002) also observed, whatever the construct, the scoring criteria of the IELTS academic writing test can clarify, at least partly, aspects of the writing construct that the test aims at testing. As the scoring criteria for the writing tests in academic and general training modules of IELTS are identical, the same underlying construct seems to be at work in both modules of IELTS (Weigle, 2002). The scoring criteria for essay writing tests are task response, coherence and cohesion, lexical resource, and grammatical range and accuracy (for descriptions of these criteria see IELTS, 2021b).

Considering the descriptions of the IELTS essay writing test and its scoring criteria together, the construct of the IELTS essay writing test can be written as a statement such as this: *The ability to write, in 40 minutes, an impromptu at least 250-word academic essay that is coherent and cohesive, has appropriate and varied lexical resources, enjoys grammatical range and accuracy, responds to the task statement/question fully and relevantly in full connected discourse without plagiarism*. If this is the construct of the IELTS essay writing test, then it seems that the test enjoys construct validity. If not, the IELTS test designers and developers should present a clear statement of what the construct of the IELTS essay writing test is. Only then can we judge and evaluate whether the IELTS essay writing test has construct validity.

To avoid the same problem, that is, a lack of a clear statement of the construct of an essay writing test, we have tried to define the construct of the WEs essay writing test used in this study as clearly as possible. To understand the construct of the essay writing test of IELTS, we have to first see what the test design and scoring criteria of the test are so that we can understand the construct of the test. However, for the WEs essay writing test, we have adopted a reverse approach in that we have tried to first see what the construct of a WEs essay writing test should be and then design the test and determine its scoring criteria in a way that best serves the purpose. To these ends, understanding the construct of a WEs academic writing test can be a first step in designing what we have named *the World Englishes Test of Academic Writing*, henceforth WETOAW (pronounced like veto). Designing WETOAW seems to be easier said than done and depends on how its construct is defined, which is another challenge in itself.

A good start can be defining the construct of a WEs test not in terms of language proficiency but *communicative proficiency*. Conventionally, proficiency is defined as language proficiency and based on native-speakerism and conformity to standard Englishes (Brown, 2014; Canagarajah, 2006). Yet, this kind of conceptualization does not seem to be suitable for a WEs test. Although language proficiency and communicative proficiency cannot be independent of each other and in testing one, the other is inevitably tested too, prioritizing *communication* over *language* means paying less attention to linguistic proficiency (Canagarajah, 2006; Jenkins, 2020) and more attention to what a person can do (Canagarajah, 2006) or how an individual can achieve a communicate outcome with whatever linguistic proficiency the person has, be it based on native speaker or nonnative speaker linguistic norms (Jenkins, 2020).

The importance of testing the ability to use a language rather than knowledge of a language (its linguistic tools) has been voiced by many scholars (see, for example, Brown, 2020; Canagarajah, 2006; Jenkins, 2020; McNamara, 2000; Tomlinson, 2010; Weir, 1990). As McNamara (2012) maintained, "current conceptualizations of proficiency in terms of gradual approximation to the competence of the native speaker will need to be drastically revised" (p. 202). This drastic revision is necessary because "there is still an insistence on 'correct' grammar and pronunciation in ELT examinations" (Jenkins, 2006, p. 43).

Yet, even if we accept to define the construct or the proficiency of a WEs test as communicative proficiency, the definition of communicative proficiency can pose its own challenges. A good start, though, can be considering Smith and Nelson's (1985) three aspects of understanding. Smith (2009) explained what these aspects of understanding refer to.

> 1. Intelligibility: the degree to which one is able to recognize a word or utterance spoken by another;
> 2. Comprehensibility: the degree to which one is able to ascertain a meaning from another's word or utterance; and
> 3. Interpretability: the degree to which one is able to perceive the intention behind another's word or utterance. (p. 17)

Smith (2009) further used some examples (a poem and a Thai utterance) to conclude that "we can have high intelligibility with low comprehensibility; and high comprehensibility with little or no interpretability" (p. 19). Although these levels of understanding were originally conceptualized for speaking, we believe that they can apply to writing too, with slight changes in what they can mean in writing (discussed later in this paper). Effective communication is then met when we understand or make ourselves understood in all these three dimensions of understanding. Other aspects of proficiency such as linguistic proficiency (i.e., grammar and vocabulary) can also be judged based on how much they contribute to communication.

Considering the essence of WEs and also Smith and Nelson's (1985) model of understanding, a definition of communicative proficiency does not seem unattainable. Excluding the fact that communication can happen without verbalized or written use of language, for example, through pantomime or facial expressions, we define communicative proficiency as this: *The ability to communicate in the sense of understanding and making oneself understood in terms of intelligibility, comprehensibility, and interpretability in native or nonnative contexts of a* *language by using native or nonnative norms or varieties of the language*. To consider English as the language for communication, the definition of communicative proficiency in English becomes this: *The ability to communicate in the sense of understanding and making oneself understood in terms of intelligibility, comprehensibility, and interpretability in native or nonnative contexts of English by using native or nonnative norms or varieties of English.*

This definition highlights that WEs considers all the existing varieties of English, whether native or nonnative. As Kachru (2013) noted, "all the users of Englishes are integral parts of the World Englishes" (p. 4). Then, appreciating different varieties of English means, when appropriate and when they have their own place in communication, native varieties should also be used. This seems the case in a test of *academic writing*. As Canagarajah (2006) noted, "in extremely formal institutional contexts where inner-circle norms are conventional (such as in academic communication), one has to adopt the established norms" (p. 234).

Therefore, although in a WEs test, and especially in its speaking and listening tests, nonnative norms should be incorporated and acceptable, in a WEs test of *academic* writing, the native norms in the sense of correct use of grammar and vocabulary should still matter because some grammatical features are more common in academic writing than in nonacademic writing or conversations (see Biber, 2006). Moreover, academic essays are usually expected to be written in correct neat grammar with rather formal vocabulary and grammar (e.g., no contractions in academic writing), whether they are written by native or nonnative students. Thus, correct grammar should matter in a WEs test of *academic* writing, though not as much as it matters in IELTS or TOEFL.

Taking the definition of *communicative proficiency* and the fact that native speaker norms have their established presence in academic communication, the construct of WETOAW, which is the ability or proficiency that we wish to test, can be defined as a statement like this: *The ability to communicate in the sense of making oneself understood in writing in terms of intelligibility, comprehensibility, and interpretability, in a formal academic style, by generating ideas, paraphrasing others' words, reasoning, critiquing, and writing pertinent content coherently and with understandable organization.*

The first half of this definition refers to communicative proficiency and the second half refers to what we believe formal academic writing entails (see also Biber, 2006), that is, the ability to generate one's own ideas, paraphrase what others have said or written, reason, critique others' ideas, know the norms of academic writing in terms of neat correct grammar and vocabulary, and the ability to write relevantly to a particular topic. Not all these skills may be tested by an essay writing test. For example, as a task of WETOAW, if we can have more than one task for it, we can present test-takers with a text and ask them to write a summary of it by paraphrasing what the author has said and critiquing his or her ideas. However, if, for practicality reasons, we have to have only one writing task, then, essay writing seems to be one of the best options as it can test many of the abilities listed in the definition of the construct of WETOAW, such as writing in a formal academic style, generating ideas, reasoning, and writing on topic (not off topic) with easy-to-follow clear organization, to name but a few. Therefore, the WEs essay writing test can be a task or the only task of WETOAW.

As far as the scoring criteria for WETOAW are concerned, it should be clear that they should be determined in a way that covers, as much as possible, different aspects of the construct of WETOAW. To do so, before proposing the scoring criteria, it is helpful to determine out of what total score the weight of scoring criteria is to be determined. In light of this, we believe that a scale of 100 is useful for scoring WEs essays for two main reasons. First, because there are usually varied criteria for scoring, a 100-point scale can allow more room and evaluation points for different criteria. Second, most people of different cultures may have a mental comfort and familiarity with percentages and, therefore, a 100-point scale may be more tangible for them than, say, a 9-point scale as in IELTS. Considering the probable advantages of a 100-point scoring scale and the definition of the construct of WETOAW, the following scoring criteria, listed tablewise, have been considered for the essay writing test of WETOAW.

Table 1
*Proposed scoring criteria for a World Englishes essay writing test*

| Communicative Proficiency | Scoring Weight | Academic Proficiency | Scoring Weight |
|---|---|---|---|
| Intelligibility | 5% | Writing in formal academic style including accurate grammar (10%) and vocabulary (10%) | 20% |
| Comprehensibility | 10% | Generating ideas, reasoning, and critiquing | 10% |
| Interpretability | 10% | Coherence (and cohesion) | 10% |
| Pertinent content | 25% | Organization | 10% |

The weight of every one of these scoring criteria can be a very subjective decision. As communication is the essence of WEs, we deemed it fair to allocate 50 percent to it. Contrary to Smith and Nelson's (1985) account of intelligibility in speaking and listening, which related mainly to prosodic features of what an individual is saying, we would like to define intelligibility in writing as the legibility of the produced text, not in the sense of the beauty of handwriting but the accuracy and understandability of spelling. Comprehensibility refers to the fact that a sentence is understandable or not, and interpretability means if test-takers have been able to make themselves and their intentions understood. We expect that all these aspects of communicative proficiency are easy to evaluate and may not put much burden on the shoulders of raters. If raters can understand test-takers' written discourse in the three aspects of understanding, they can easily give the 25 percent scoring points allocated to them (see Table 1).

However, all the three aspects of understanding can be present, yet a text may be quite off-topic, not addressing the task instruction or topic. Thus, pertinent content is vitally important because individuals should be able not only to communicate (in the three aspects) but also to write to the point and relevant to the topic under discussion, as it is expected in academic writing. Therefore, providing pertinent content can take 25 percent of scoring because it can be as important as communicating in the basic three aspects of understanding. It may, then, not be unfair to allocate at least half of the scoring (50%) to communication in the sense of making oneself understood intelligibly, comprehensibly, interpretably, and pertinently.

The second half of scoring is related to what usually matters in academic writing. Accurate grammar and vocabulary each can take 10 percent of the scoring. Generating ideas, reasoning, and critiquing all can take 10 percent, which may not seem much. They are not given much weight in scoring because although they are important, judging them can be too subjective, susceptible to human error, and therefore requires a lot of training on the part of raters. Coherence can take 10 percent, which includes cohesion too. That is, a separate score for cohesion has not been considered because cohesion can be considered a tool or means for achieving coherence. And finally, organization has been separated from coherence because coherence is usual even within one paragraph, but organization is usually judged based on all the paragraphs in an essay. All in all, then, by allocating half of the scoring of an essay to *communicative proficiency* and the other half to *academic proficiency*, we may be able to have a fairer scoring policy than what is currently exercised in IELTS.

Inevitably, determining the scoring criteria of a writing test is mostly subjective and our proposed criteria are not exceptions as they are based on our own understanding of what matters in academic writing in WEs contexts. We, therefore, welcome criticisms of the scoring criteria and hope that our proposed WEs scoring criteria are thought-provoking starting points for further proposals as to how to define the construct of and proficiency in WETOAW and how to score such a test accordingly.

## 3. Method

### *3.1.Participants and Setting*

The participants of this study were 17 Iranian IELTS candidates (7 males, 10 females, age mean 28) and three Iranian IELTS instructors (2 males, 1 female, age mean 35), who were the raters of the essays.

The IELTS candidates were of different language learning backgrounds. They had been learning English for some years, and they had experience in studying for IELTS for one year. They all were Persian native speakers who had learned English as a foreign language. The candidates had hands-on experience of the IELTS academic writing test, experiencing the test conditions in their IELTS preparation courses, in some mock tests, and in the IELTS test. Twelve of them had taken the IELTS academic writing test and five of them had taken the IELTS general training test. And, they were selected based on purposeful, convenience, and snowball sampling methods.

The instructors or raters had experience in teaching English for 10 years and teaching for IELTS for 5 years on average. They all were Persian native speakers who had learned English as a foreign language for many years as well as in their B.A. and M.A. programs, and they all were Ph.D. students of applied linguistics at Kharazmi University. All the three instructors had already taken the real IELTS test (with overall band scores of 7.5, 8, and 8), and all of them were deeply involved with IELTS, having both theoretical and practical knowledge of it. The instructors were selected based on convenience and purposeful sampling methods. It was purposeful sampling because we looked for the raters who had taken the real IELTS test, who had been teaching it for some years, and who had both theoretical and practical knowledge of it. And it was convenience sampling because we used only the available instructors who had the necessary experiences.

### *3.2.Instrumentation*

To collect participants' essays, two sets of essay topics were written and used as the instrument of the study. In determining some essay topics for the study, it was important to choose topics that could be of interest to most of the participants and that were relevant to their lives. This was done in an attempt to increase the authenticity of the topics (see Bachman & Palmer, 1996; Behizadeh & Engelhard, 2014). After selecting the topics, we wrote them twice, once for the IELTS essay writing test and once for the WEs essay writing test. The WEs topics were the same as the IELTS topics, but the WEs topics had more explanations as to what candidates are expected to do. Overall, six topics were formed, three in the form of agree/disagree topics and three in the form of direct opinion-seeking questions (the topics can be seen in the Figshare data repository with the doi address of https://doi.org/10.6084/m9.figshare.19817452).

Expert opinion was sought to check whether the essay topics are valid instruments for obtaining data, and the interrater reliability of the scores given by the three raters was estimated, showing that the Pearson correlation coefficients were more than .90 in all cases (the reliability estimates can be seen in the Figshare data repository with the doi address of https://doi.org/10.6084/m9.figshare.19817452).

As far as the training of the raters is concerned, to train the raters and make them more familiar with scoring IELTS-wise, IELTS scores guide was used (see IELTS, 2018), and to train them for the WEs essays, the proposed scoring criteria for the WEs essay writing test were used (see Table 1).

### *3.3.Procedures*

The first stage of data collection was writing some essay topics, as explained in the previous section. After selecting the candidates, we asked them to read the topics and choose one pair of them. Then, they were asked to write two academic essays about their chosen topic, one under the IELTS test conditions, a 250-word essay in a 40-minute uninterrupted time frame, and one under the WEs test condition. In determining the WEs test condition, we tried to focus more on the authenticity aspect of test usefulness by giving candidates 120 minutes to write an essay, without determining any minimum word requirement. And, there was a two-week gap between the IELTS essay and the WEs essay.

After all the essays were written, we applied four scoring scenarios. In the first and second scenarios, both IELTS essays and WEs essays were scored based on the IELTS scoring criteria. In the third and fourth scenarios, both IELTS essays and WEs essays were scored based on the WEs scoring criteria (see Table 1).

All the essays were scored by the three raters once, independently of each other. The raters were also cautioned about the halo effect and were briefed on the tenets of WEs. Moreover, samples of IELTS essays of each band and half band score (taken from IELTS, 2018) were reviewed and discussed with the raters to make them more familiar with scoring IELTS-wise. The same was done with the same samples of IELTS essays but scored based on WEs criteria. The WEs essays of the participants of this study were not used for the briefing sessions so that raters' scoring of the essays written for this study remain independent of each other (for the importance of training raters see Doosti & Ahmadi Safa, 2021; Fahim & Bijani, 2011; Ghanbari & Barati, 2014).

It is also to be noted that the study adopted a quasi-experimental design to investigate whether there is a statistically significant difference among Iranian IELTS candidates' obtained grades in essay writing tests when the essays are written under IELTS and World Englishes test designs and scored by IELTS and World Englishes scoring criteria. In what follows, the results of the scoring are presented and discussed.

## 4. Results

As mentioned earlier, each essay was scored by adopting IELTS and WEs scoring criteria. The mean scores of the three raters can be seen in Table 2.

Table 2
*Scoring Results: Mean Scores of the 3 Raters*

| Test taker | IELTS Essay Scored By IELTS Scoring Criteria | WEs Essay Scored By IELTS Scoring Criteria | IELTS Essay Scored By WEs Scoring Criteria | WEs Essay Scored By WEs Scoring Criteria |
|---|---|---|---|---|
| No. 1 | 7 | 8 | 8 | 8.5 |
| No. 2 | 5 | 6 | 6.5 | 6.5 |
| No. 3 | 6 | 7 | 7.5 | 8 |
| No. 4 | 6 | 8.5 | 8 | 8.5 |
| No. 5 | 6.5 | 8 | 7.5 | 8.5 |
| No. 6 | 6.5 | 8 | 8 | 8 |
| No. 7 | 8.5 | 8.5 | 8.5 | 9 |
| No. 8 | 7.5 | 8.5 | 8.5 | 9 |
| No. 9 | 5 | 7 | 7 | 7.5 |
| No. 10 | 5.5 | 6 | 7 | 7 |
| No. 11 | 4.5 | 6.5 | 6.5 | 6.5 |
| No. 12 | 8 | 8.5 | 8.5 | 9 |
| No. 13 | 6.5 | 8 | 8.5 | 9 |
| No. 14 | 7.5 | 8.5 | 8.5 | 9 |
| No. 15 | 8 | 8.5 | 8.5 | 9 |
| No. 16 | 7.5 | 8 | 8 | 8.5 |
| No. 17 | 4.5 | 7.5 | 7 | 8 |

*Note*. WEs scoring criteria were designed in a way that essays were scored out of 100. These scores were then proportioned to 9 so that they could be compared with IELTS scores.

As mentioned earlier, the interrater reliability estimates of the scores given by the three raters in the above-mentioned four scoring scenarios obtained through the Pearson correlation coefficient were more than .90 in all cases. After obtaining the essay grades (the scores), a two-way repeated-measures ANOVA was run, whose results are as follows. Let us begin with descriptive statistics in Table 3.

Table 3
*Descriptive Statistics of the Repeated-Measures ANOVA*

| *Descriptive Statistics* | | | |
|---|---|---|---|
| | Mean | Std. Deviation | N |
| IE_ISC | 6.4706 | 1.26825 | 17 |
| WEE_ISC | 7.7059 | .88492 | 17 |
| IE_WESC | 7.7647 | .73139 | 17 |
| WEE_WESC | 8.2059 | .86709 | 17 |

*Note*. IE_ISC stands for IELTS essays scored by IELTS scoring criteria. WEE_ISC stands for Wes essays scored by IELTS scoring criteria. IE_WESC stands for IELTS essays scored by Wes scoring criteria, and WEE_WESC stands for Wes essays scored by Wes scoring criteria.

As can be seen in Table 3, participants' essay grades were lowest when both the test design and the scoring criteria were based on IELTS (mean = 6.47). When the WEs essays were scored by IELTS scoring criteria, the mean was 7.70, which is similar to the mean of essay grades when IELTS essays were scored by WEs scoring criteria (mean = 7.76). Finally, the highest mean was obtained when essays were written under WEs test design and scored by WEs scoring criteria, with a mean of 8.20. Let us now see the results of the repeated-measures ANOVA in Table 4.

Table 4
*Results of the Repeated-Measures ANOVA*

| | df | F | Sig. | ηp2 |
|---|---|---|---|---|
| Test Design | 1 | 102.648 | .000 | .865 |
| Scoring Criteria | 1 | 56.751 | .000 | .780 |
| Test Design - Scoring Criteria | 1 | 18.000 | .001 | .529 |

As can be seen in Table 4, there was a statistically significant main effect of test design on participants' essay grades (F(1, 16) = 102.64, p = .000, ηp2 = .865). Following Cohen's (1988) guideline as to how to determine the magnitude of effect sizes, we can see that the Partial Eta Squared (*ηp2*) or effect size here is .865, which is much more than what is conventionally considered a large effect size (i.e., 0.14). This shows that test design has had a huge impact on participants' essay grades. Similarly, there was a statistically significant main effect of scoring criteria on participants' essay grades (F(1,16) =56.751, p = .000, *ηp2* = .780), with an effect size of .780, showing that scoring criteria have had a big impact on participants' essay grades. And finally, there was a significant interaction between test design and scoring criteria (F(1,16) = 18.00, p = .001, *ηp2* = .529), with an effect size of .529, showing that the interaction between test design and scoring criteria is significant.

The above information answers the research question of this study, showing that there is a statistically significant difference among Iranian IELTS candidates' obtained grades in essay writing tests when the essays are written under IELTS and WEs test designs and scored by IELTS and WEs scoring criteria. Let us see what the results mean from a more practical point of view.

## 5. Discussion

What all the results suggest is that increasing the authenticity of a test may better show test-takers' highest possible performance in test conditions and by extension may better represent their performance in non-test conditions. This is so because "the way test takers perceive the relative authenticity of test tasks can, potentially, facilitate their test performance" (Bachman & Palmer, 1996, p. 39), and also because increasing authenticity means increasing the generalizability of test-takers' performance on test tasks to nontest tasks. Moreover, increasing authenticity will increase the construct validity of a test too. Considering authenticity as an important part of construct validation, Bachman and Palmer (1996) defined authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU (target language use) task" (p. 23). This, then, means that, as mentioned above, performance on a test task should be generalizable to and representative of the performance on nontest tasks.

Yet, this does not seem to be the case in the IELTS essay writing test because in real nontest settings, few people may write a 250-word essay in 40 minutes and, therefore, the IELTS essay writing test does not seem to enjoy authenticity much. There may not be much point in testing the writing ability of test-takers when a writing test is far from real-life writing conditions, or technically put, the TLU domain. Therefore, at least as long as the authenticity of the IELTS essay writing test is concerned, it seems that there is room for improvement.

The findings of this study suggest two solutions for IELTS candidates' problem of scoring low in writing (see the Introduction). The first solution is to make an academic writing test as authentic as possible for test-takers who wish to communicate in international contexts. To do so, the *test design* can be more WEs-based, which can be done by simulating, as much as possible, the characteristics of writing in the target language use domain. And, the easiest way to do so in the IELTS essay writing test seems to be increasing the time allotment of the test.

Although increasing the test time may seem in conflict with the practicality of administrating a test, it may not burden the administrators of a test much and it may not incur considerable costs if a 40-minute essay writing test is changed into, for example, a 60-minute, 90-minute, or even 120-minute test. Even if it is burdensome and costly to increase the time of an essay writing test, it may be necessary to do so because test-takers are entitled to be tested in a way that best allows them to show their writing ability.

Moreover, if a test is to be considered international, the test design should be as international and real-life as possible. This is much more important in a high-stakes test such as IELTS because many test-takers may not score well simply because the time allotment seems to be quite unrealistic as long as real-life conditions are concerned. It is quite possible that if test-takers have more time for the essay writing test, they will be able to write noticeably better essays and score better in essay writing tests, a case that has been supported by the results of this study.

The second solution for the problem of IELTS candidates scoring low in writing may be changing the scoring criteria of the test. Since the four scoring criteria of the IELTS essay writing test are equally weighted, half of the scoring (50%) has been allocated to linguistic proficiency in the sense of correct and varied use of grammar and vocabulary. We believe that this is too much scoring weight to be allocated to an essay written in WEs contexts, in which communication matters most, and correct grammar and vocabulary, though still important in academic writing, may not take *half* of the scoring. We also believe that if our proposed scoring criteria for a WEs essay writing test (see Table 1) are also used for scoring IELTS essays, then WEs scoring criteria may be the second solution to the IELTS candidates' problem of scoring low in writing.

## 6. Conclusion

The main conclusion that may be drawn from the results of this study is that the reasons why IELTS candidates usually score low in writing may be rooted, among other things, in the *test design* and *scoring criteria* of the IELTS essay writing test, not necessarily in IELTS candidates' weaknesses in writing, though the latter issue can be the case too.

As long as *test design* is concerned, the participants of this study could write better essays under the WEs test condition, suggesting that they were not as weak in writing as the essays written under the IELTS test condition may suggest. Since the main difference between IELTS and WEs test designs was increased authenticity of the WEs test (i.e., increased test time and clearer explanations in the instructions of the WEs essay topics), it may, then, follow that increasing the authenticity of the IELTS essay writing test in favor of WEs contexts may go a long way in solving the IELTS candidates' problem of scoring low in writing.

The second conclusion has to do with the scoring criteria. It was shown that the participants' IELTS essays could be scored significantly higher just when WEs scoring criteria were applied to them. This may, thus, be a second solution to the IELTS candidates' problem of scoring low in writing, and this solution seems to be more practical than changing the test design. This is so because as long as the administration of the IELTS writing test is concerned, increasing the time allotment may be difficult from a practical perspective. However, changing the scoring criteria may be less troublesome as IELTS raters all go through extensive training, and once trained well, scoring IELTS essays based on the new WEs scoring criteria may not be much more (or any more) time-consuming for them. Overall, then, there seem to be at least two solutions to the IELTS candidates' problem of scoring low in writing, which are applying the WEs test design and WEs scoring criteria.

The last conclusion we wish to draw relates to incorporating WEs in language assessment. In light of this, the literature of WEs (see, for example, Brown, 2014) suggests that assessing test-takers' performance in WEs contexts is so challenging that there have been few practical attempts to design any test of WEs. One reason for this may be an idealistic attempt to solve all the theoretical problems of incorporating WEs into assessing language practices (see Hu, 2021). However, delaying the development of a WEs test because of an endeavor to form a perfect or almost perfect conceptualization of how WEs can be incorporated into language assessment may mean developing a WEs test too late. This is so because not only will we never be able to have an agreed-upon way of testing WEs but also the current high-stakes large tests such as IELTS and TOEFL will become, if have not already become, so strong as to form a complete testing hegemony or imperialism that is extremely difficult, if not impossible, to compete with. As Khan (2009) discussed, "power and control exerted by a dominant class in society is sustainable if it gains support through 'consent' of the masses" (p. 191). Therefore, as long as native-speaker-based tests such as IELTS and TOEFL gain support from the public, it can be difficult for a WEs test to stand out. And, why should the ones who have power give it away? As Jenkins (2020) put it, "the large international testing bodies are unlikely to relinquish their key and highly remunerative role in university English language entry testing unless/until they see an equally key and highly remunerative role for themselves in adopting an ELF approach" (p. 6). Yet, what is difficult is not necessarily impossible, and it is hoped that this study, limited as it may be, can shed some light on how a WEs test may compete with the existing large tests (such as IELTS) on a much more equal footing. To this end, a move forward can be the replication of this study when its limitations have been answered.

Concerning the limitations of the study, some points are also in order. The main limitation of the study was its small sample size, which may not warrant the above-discussed conclusions. Yet, despite recruiting a cohort of Iranian IELTS test-takers, the results of the study can potentially apply to wider groups of IELTS test-takers who have difficulty with the IELTS essay writing test. This calls for further research to see if the results of this study will be the same if the study is replicated in other contexts. Another limitation of the study was a lack of intrarater reliability estimates. The raters had accepted to cooperate in the study only if they had to score each essay once based on IELTS and once based on WEs scoring criteria. They could not be asked to rate each essay twice based on the different criteria. Therefore, it was not possible to have intrarater reliability estimates.

Yet, with all its limitations, we hope that this study, however preliminary it may seem, can be considered as a practical attempt to contribute to the development of a WEs test by designing, defining the construct, and determining the scoring criteria of what we have named the World Englishes Test of Academic Writing (WETOAW). We believe that, compared to the currently used large-scale tests such as IELTS, WETOAW may better show test-takers' true writing skills. This is so because the results of this study showed that the participants could score highest when both WEs test design and scoring criteria were applied. In other words, compared to IELTS, WETOAW may be a better means of showing test-takers' best performance in essay writing, as shown in this study.

This study and its results can be important for IELTS candidates, international students, policy-makers, language test developers, WEs and ELF researchers, and testing experts, to name but a few. We see it quite possible that if WETOAW is supported and promoted and then accepted by universities and organizations around the world, those who wish to study in international contexts may have better options than taking IELTS or similar tests (such as TOEFL). We hope that this study can also show that the future of assessing writing in WEs contexts is not gloomy and that it is not impossible to test writing, even academic writing, in WEs contexts.

**Acknowledgments**
We wish to thank all the participants and raters of the study for their cooperation.

**Declaration of Conflicting Interests**
There are no conflicting interests in this strudy.

**Funding**
There is no funding to report for this study.

## References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Behizadeh, N., & Engelhard Jr, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing*, *21,* 18-36.

Biber, D. (2006). *University language: A Corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.23

Brown, J. D. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly*, *11*(1), 5–26. https://doi.org/10.1080/15434303.2013.869817

Brown, J. D. (2020). World Englishes and international standardized English proficiency tests. In C. Nelson, Z. Proshina, & D. Davis (Eds.), *The handbook of world Englishes* (2nd ed., pp. 703–724). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119147282.ch39

Canagarajah, S. (2006). Changing communicative needs: Revised assessment objectives, testing English as an international language. *Language Assessment Quarterly*, *3*(6), 229–242. https://doi.org/10.1207/s15434311laq0303_1

Cohen, J. (1988). *Statistic power analysis for the behavioral science* (2nd ed.). New York: Academy Press.

Doosti, M., & Ahmadi Safa, M. (2021). Fairness in Oral Language Assessment: Training Raters and Considering Examinees' Expectations. *International Journal of Language Testing*, *11*(2), 64-90.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters’ severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Ghanbari, N., & Barati, H. (2014). Iranian EFL writing assessment: The agency of rater or rating scale?. *International Journal of Language Testing*, *4*(2), 204-228.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge: Cambridge University Press.

Hu, G. (2021). Language assessment in global Englishes. In A. F. Selvi & B. Yazan (Eds.), *Language teacher education for global Englishes: A practical resource book* (pp.199–206). London: Routledge.

IELTS (2018). IELTS scores guide. Retrieved 22, 03, 2021, from https://ielts.kz/wp-content/uploads/2019/01/ielts_score-guide_a4_2018_web.pdf

IELTS (2021a). For research, test statistics, test taker performance 2019, band score information, 2019. Retrieved 11, 12, 2021, from https://www.ielts.org/research/test-taker-performance.

IELTS (2021b). For test-takers, test format. Retrieved 8, 10, 2021, from https://www.ielts.org/for-test-takers/test-format#tab-4

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal*, *60*(1), 42-50. https://doi.org/10.1093/elt/cci080

Jenkins, J. (2020). Where are we with ELF and language testing? An opinion piece. *ELT Journal*, *74*(4), 473-479.

Kachru, B. B. (2013). History of world Englishes. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–12). London: Blackwell Publishing Ltd.

Khan, S. Z. (2009). Imperialism of international tests: An EIL perspective. In F. Sharifian (Ed.), *English as an international language: Perspectives and pedagogical issues* (pp. 190–205). Bristol, UK: Multilingual Matters.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

McNamara, T. (2012). English as a lingua franca: The challenge for language testing. *Journal of English as a Lingua Franca*, *1*(1), 199–202. https://doi.org/10.1515/jelf-2012-0013

Pearson, W. S. (2019). Remark or retake? A study of candidate performance in IELTS and perceptions towards test failure. *Language Testing in Asia*, *9*(1), 1-20.

Read, J. (2015). *Assessing English proficiency for university study*. Hampshire: Palgrave Macmillan.

Smith, L. E. (2009). Dimensions of understanding in cross-cultural communication. In K. Murata & J. Jenkins (Eds.), *Global Englishes in Asian contexts: Current and future debates* (pp. 17–25). Basingstoke, UK: Palgrave Macmillan. https://doi.org/10.1057/9780230239531_2

Smith, L. E., & Nelson, C. (1985). International intelligibility of English: Directions and resources. *World Englishes, 4,* 333–342. https://doi.org/10.1111/j.1467-971x.1985.tb00423.x

Tomlinson, B. (2010). Which test of which English and why. In A. Kirkpatrick (Ed.), *The Routledge handbook of world Englishes* (pp. 599–616). London: Routledge. https://doi.org/10.4324/9781003128755-44

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. (1990). *Communicative Language Testing*. Hemel Hempstead: Prentice-Hall.