

The International English Language Testing System (IELTS): A Critical Review

Peter Peltekov

University of Calgary, Canada
e-mail: pnpeltek@ucalgary.ca

Abstract

Considering the increasing popularity of the International English Language Testing System (IELTS), the present article provides a succinct description and critique of the test. As with any high-stakes assessment, educational institutions need to carefully examine all aspects of a given assessment tool before applying it in practice. Green's (2014) framework for the evaluation of second language assessment tools was applied to the analysis of the IELTS test. The present review demonstrated that there are many ways in which the IELTS test can be improved (e.g., increasing the authenticity of the listening modules and reducing the role of construct irrelevant skills). While it is far from flawless and not the only option, IELTS continues to be one of the most popular international tests of English language proficiency. Clearly, the test is an important gate-keeping measure and an incentive for millions of non-native speakers to improve their English language skills. As we know, the beneficial consequences of a given assessment system are on the top of the hierarchy of effective assessment characteristics (Green, 2014), and IELTS seems to achieve its purpose. However, it is hoped that the present critical review is a valuable contribution to the ongoing validation and improvement of the test. At the very least, it is hoped that it would help assessment stakeholders to better understand the structure of the test and to reflect on its usefulness in a more informed and objective way.

Keywords: IELTS, language testing, assessment characteristics, test reliability, test validity.

1. INTRODUCTION

The International English Language Testing System (IELTS) is a large-scale test designed to assess the English language proficiency of non-native English speakers willing to study or work in places where English is the language of communication. Established in 1989 and jointly owned by the British Council, IDP: IELTS Australia and Cambridge Assessment English, IELTS is now an international test and it is growing in popularity. At present it is recognized by more than 10,000 educational institutions in over 140 countries (IELTS, 2020). In the UK, for example, it is one of the most widely recognized tests of English for academic purposes (Green, 2007a). Clearly, IELTS is a very popular tool for assessing English language abilities. However, while its popularity can hint at the general usefulness of the test, it should not be automatically assumed that the test is equally effective in all settings. Moreover, the test has undergone substantial modifications since it was first created. While there are existing studies on the effectiveness of the test, it is important that researchers in language testing continue to evaluate the usefulness of the test as it evolves. The aim of the present study is to provide a description and a critical review of the most recent version of the IELTS test. The outcomes of this analysis have important implications for stake-holders, such as educational institutions using, or planning to use, the IELTS test as a test of academic English abilities.

2. THE QUALITIES OF USEFUL ASSESSMENT

When analysing a given assessment tool, it is important to select an appropriate set of criteria for the description and evaluation of the test. Green (2014) provides a framework for the evaluation of assessment systems. In his model, there are four main qualities of useful assessments: practicality, reliability, validity, and beneficial consequences, as shown in Figure 1 below.



Figure 1. *The main qualities of useful assessment.*

The form of a cone conveys the relationship between the different qualities. At the base we see *practicality*, which is an essential prerequisite for all assessment. If a test is not practical and it cannot be easily implemented, then the presence of the other qualities would be irrelevant. On the top of the cone, we see the ultimate goal of any assessment system: the *beneficial consequences* (e.g., providing accurate measures of language skills that can be used for decision-making purposes). In the middle there are two of the central qualities of effective language tools: *reliability* (i.e., to what extent the test scores are consistent) and *validity* (i.e., to what extent the test measures what it is supposed to measure). According to Green (2014), other important qualities, such as authenticity or absence of bias, are well-accounted within

this four-quality framework. This model was used in the review of the IELTS test, and the results of the analysis are discussed in relation to the four main qualities of effective assessment.

2.1 Description of the IELTS test

Before discussing the quality of the test, it is important to summarize how exactly the IELTS test is structured. For those who are not (sufficiently) familiar with the test, IELTS offers an official online portal with rich information about the test structure in the form of handbooks for institutions, teachers, and students, in addition to sample practice test materials (IELTS, 2020). The test consists of four separate modules (listening, reading, writing, and speaking), and it is available in two versions: IELTS Academic and IELTS General Training. The Academic test is designed for individuals applying to study in postsecondary institutions, and it measures English language proficiency needed for the academic environment. The General Training test, on the other hand, is for people willing to migrate to an English-speaking country, and it measures English language proficiency in practical workplace and social contexts. Test results are reported as scores on a nine-band scale from “non-user” (band 1) to “expert user” (band 9).

2.2 Characteristics of the Rubric

The critical evaluation of a language test should begin with a closer analysis of the rubric – a term that Douglas (2000) uses to refer to the information about the test structure and specifically about “how test takers are expected to proceed in taking the test” (Bachman, 1990, as cited in Douglas, 2000). The first important element of the test specifications is the statement of test purpose (Green, 2014), also called specification of objective (Douglas, 2000). The main objective of IELTS is to provide evidence of English language proficiency, which often serves as a gate-keeping measure. In addition, the test is also used “to guide decisions about the amount of language study required for students to satisfy admissions requirements” (Green, 2005). Green (2005) warns, however, that gate-keeping decisions should be based on multiple sources of evidence.

The rubric also provides test takers with information about the structure of the test, the time allotment for each task, and the procedures that are followed. This information is summarized in Table 1.

Table 1. Structure of the IELTS test

SECTION	TIME	QUESTIONS	TASKS	EVALUATION
Listening	30 min	40 (10 per section)	A variety of task types are used. Tasks are divided in 4 sections: 1) conversation on everyday topics 2) monologue on everyday topics 3) conversation on academic topics 4) monologue on academic topics	Each question is worth 1 mark

Reading	60 min	40	<u>Academic version:</u> - 3 reading passages on academic topics (at least 1 has a logical argument)	<u>General version:</u> - several shorter texts divided in 3 topics: 1) social survival 2) workplace survival 3) general reading	Each question is worth 1 mark
			Task types are identical for both versions.		
Writing	60 min	2	<u>Academic version:</u> 1) describe visual information (e.g., graph) 2) respond to a point of view or argument	<u>General version:</u> 1) Respond to a situation (e.g., in a letter format) 2) respond to a point of view or argument	Band scores awarded based on performance descriptors
Speaking	11-14 min	3	There are 3 tasks: 1) introduction and general questions 2) monologue on a given topic 3) discussion related to 2)		Band scores awarded based on performance descriptors

As shown in Table 1, the test starts with the Listening module, which lasts approximately 30 minutes. Test takers are instructed to write their answers on the question paper as they listen to the recordings, and at the end they are given additional 10 minutes to transfer their answers to the official answer sheet. It is important to note that for this module there is no distinction between academic and general version, meaning that all test takers listen to recordings on both general and academic topics. There are 40 questions, and the question types include: multiple choice, matching, plan/map/diagram labelling, form/note/table/flow-chart/summary completion, and sentence completion (IELTS, 2020). Evaluation is straightforward – each correct answer is awarded one point.

The Reading module is twice longer than the listening module, and test takers are not allowed extra time to transfer their answers to an answer sheet. First of all, asking assesseees to copy their answers from a question paper to an answer sheet is a complexity that should be avoided, as it can lead to increased opportunities for errors to occur (Green, 2014). Moreover, it is especially plausible that test takers will commit spelling mistakes while transferring their answers if they are under time pressure. Finally, grammar and spelling errors on the answer sheet are penalized on the first two modules (IELTS, 2020), and this is a methodological issue that will be discussed in the next subsection.

Unlike the Listening module, the Reading module differentiates between an academic and a general version. Texts used in the Academic Reading test originate mostly from books and journals and may include non-verbal materials such as graphs and diagrams. The General Reading test, on the other hand, includes texts from common everyday life or work-related sources such as job advertisements, training manuals, newspapers, etc. In addition to all the task types used in the Listening module, the Reading module (in either version) can also ask test takers to identify the writer's views/claims and to provide short written answers. Scoring

criteria are similar to the ones used in the Listening module – each question is worth one mark and the maximum score is 40 points.

The Writing module has the same length as the Reading module – 1 hour. The first task requires a shorter response (~150 words) in both the Academic and the General Writing module. Therefore, test-takers are advised to spend more time on the second task, which is longer (~250 words) and contributes twice as much to the Writing band score. Responses on both tasks are assessed on four criteria: task achievement/completion, coherence and cohesion, lexical resource, and grammatical range and accuracy. The rubric also warns that instances of plagiarism are severely penalized. It is not explained, however, whether test-takers are allowed to use other sources, and if not, how plagiarism can possibly occur during the test. We can assume that plagiarism in such case would imply copying materials from other parts of the test, but not all students may be equally aware of the different forms of plagiarism. As Douglas explain, evaluation criteria should be explicitly “spelled out for the test takers to ensure that they are all equally aware of them as they plan their responses to the test tasks” (p. 53).

The last part of the IELTS test is the speaking module. It is the shortest module (11-14 minutes), and it can take place in a different day of the week before or after the other test components are covered. There are three separate and equally important tasks, all of which are recorded: 1) introduction and interview (a general conversation between the assessor and the assessee lasting for up to 5 minutes), 2) long turn (an individual talk on a given topic for approximately 3-4 minutes, including preparation time), and 3) discussion (a 4-5-minute-long conversation about the issues addressed in the second task). There are four assessment criteria used in this module: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. The speaking band score is the average of the four criteria, which are weighted equally.

3. CRITIQUE OF THE FOUR MODULES

In this section, the input and the expected response will be discussed in more detail for each of the four modules. Moreover, it will be examined how the input and the response interact and how they can affect some of the qualities of useful assessment proposed by Green (2014): practicality, reliability, validity, and beneficial consequences.

3.1 Listening

The input data in the listening module consists entirely of scripted (i.e., non-authentic) materials. This feature reduces the level of authenticity, which is one of the most important characteristics of effective assessment tools (Bachman & Palmer, 1996). The lack of significant correlation between scores on the listening module and subsequent academic performance (Aryadoust, 2012) can therefore be partially attributed to the limited situational authenticity. It is expectable that if the input does not reflect the target language use (TLU) domain closely enough, the predicative validity of the assessment will be low, which would also reduce the beneficial consequences of the test. The non-reciprocal relationship between the input and the responses (i.e., the input in the recordings cannot be altered by the responses) can also affect the situational authenticity. Although non-reciprocal test tasks are a necessary component of language tests (Douglas, 2000), the lack of reciprocal tasks may lead to

underrepresentation of the listening construct, as in many TLU situations speakers can influence the input through interpersonal interaction.

Another potential issue in the listening module's design is the need for construct-irrelevant skills (e.g., reading and writing) to demonstrate listening comprehension. As Green (2014) explains, a major problem in the assessment of receptive skills is that they cannot be directly observed. Instead of asking test-takers to read the answer choices, however, they could be allowed to listen to them as in a listening comprehension test there should be "no reading involved either in the question prompts or in the answer choices" (Vandergrift, 2006). Moreover, poor spelling and grammar in the answer sheet should not be penalized as long as the responses are intelligible. While the TLU domain often requires integration of different skills, the effects of construct-irrelevant skills should be minimized in order to improve the validity of each of the test modules.

Although multiple-choice questions (MCQs) may be a good way to assess receptive skills, providing just a few possible answers can encourage guessing and distort the results. Each of the MCQs in the listening module offers only three answer choices, meaning that test takers have over 33% chance of guessing the correct answer. As Aryadoust (2012) demonstrates, "[l]ow-ability people who have received training in test-taking strategies appear to be taking advantage of this fact, leading to flawed test results" (p. 56). Thus, it can be argued that this test task needs improvement.

Most of the recordings are short and responses normally refer to a specific part of the audio input, meaning that the scope of the relationship between the input and the response is rather narrow. Moreover, for most tasks, test takers are given time to read the questions before they listen to the recording, which draws their attention only to specific details. As Green (2014) suggests, however, listening tests should assess a variety of listening types such as listening for gist and interactive listening. On the other hand, allowing students to preview the questions may reduce some memory-related issues, and it can improve their listening comprehension performance (Hemmati & Ghaderi, 2014).

Finally, no background knowledge is required by test takers, as responses are entirely based on the input. Despite the directness of the relationship between input and response, topic knowledge is among the factors that can influence the comprehension process (Green, 2014). A review of the sample tasks shows that there is a great range of topics included, which is a proof of effective content sampling.

3.2 Reading

The reading module shares many common features with the listening module (e.g., similar procedures for responding, similar response types, and similar input-response interactions), and a lot of the points discussed above can be applied to this module as well. A major difference is that the input data for the reading module is visual and considerably longer – test takers need to read texts with a total length of 2,150-2,750 words and complete the related tasks in approximately 60 minutes (IELTS, 2020). Another difference is that the input consists of authentic readings, meaning that the level of situational authenticity is high. In the academic version, for example, assesses' reading comprehension is assessed on texts from

sources that they might be using in their later academic life. No specific knowledge is required, and glossaries are provided for technical terms.

The general version of the reading module is also intended to approximate real-life language use. For example, section two of the General Training version has been modified to focus more on a work context (e.g., work policies, applying for jobs, pay and conditions), and this change has been made as a result of the feedback that earlier versions of IELTS have received (Green, 2007b). This is an example of how the test developers are striving for continuous improvement, and how test trialing and validation are an important part of the assessment cycle (Green, 2014).

Besides the task types used in the listening module, the reading module includes tasks that involve recognising writers' opinions, attitudes, and purpose. Such types of tasks are consistent with the task-based approach where some tasks require higher degrees of comprehension (Pawlikowska-Smith et al., 2013).

The expected responses involve minimal written production (e.g., no more than two words in the completion question types). In this aspect, validity is not threatened, but poor grammar and spelling are penalized as in the listening module. Moreover, answers that do not follow the response procedures (e.g., answering in more than two words) are not accepted even if they demonstrate reading comprehension. Clearly, construct irrelevant skills (e.g., test-taking skills) play a critical role in this test module as well.

The level of reactivity and directness are identical to those of the listening module. Some tasks in the reading module, however, are found to be broader in scope (e.g., the matching task types, which require test takers to make connections between different parts of the whole text). This is an important feature as it allows different types of reading skills to be tested (Green, 2014).

3.3 Writing

In both the academic and the general version of the writing module, more weight is given to the second task. In this task, test takers are required to write an essay of approximately 250 words in response to a point of view or an argument. In both versions, the response is based on personal experiences and general knowledge. The sample writing tasks published on the official IELTS online portal (IELTS, 2020) show that the academic writing topics (e.g., Forms of transportation) are not much more formal than the general writing topics (e.g., Funding for retirement homes). The academic writing Task 2 has been criticized in previous research for being more representative of non-academic genres (e.g., letter to the editor) and spoken discourse than of the type of writing required at the university (Cooper, 2013; Moore & Morton, 2005). The main argument behind such claims is that test takers base their response on personal opinions rather than on empirical evidence. As Green (2007b) explains, however, since the 1995 revision the IELTS test does not require any specific background knowledge, while there is still a sufficient overlap between the construct of academic writing and how it is represented in IELTS. Although the level of authenticity is reduced when the response is more opinion-based and more informal compared to the typical academic expectations, requiring field specific knowledge from the test takers would threaten the validity of the writing module, which is meant to assess writing skills and not subject-related knowledge.

Task 1 of the writing module is more different between the two versions. The general version requires test takers to respond to a common everyday situation, and one of the examples provided in the handbook for candidates is: “writing to a local newspaper about a plan to develop a local airport” (IELTS, 2020). It is questionable how plausible or common such a situation would be in the TLU domain. The academic version, on the other hand requires test takers to interpret information in the form of a graph, table, chart or diagram. The relationship between the input and the response in this type of task is broader and more direct compared to Task 1 in the general version, as test takers are expected to base their response entirely on the input.

Alavi and Masjedlou (2017) found that test takers and instructors perceive the first writing task in the academic version to be intelligence-based and to require more time for thinking than for writing. Students’ underperformance on this cognitively demanding task can emotionally affect their performance on the following task (Alavi and Masjedlou, 2017). Compared to Task 2, the critique of Task 1 is reversed – we have a higher level of authenticity (i.e., it is highly plausible that real-life academic tasks would involve more thinking than writing) and a lower level of validity (e.g., diagram reading is construct-irrelevant for assessment of writing skills). Therefore, we can conclude that the presence of two different tasks is beneficial, as the potentially negative effects of the two tasks are counterbalanced.

3.4 Speaking

The speaking module includes three different tasks (see Table 1), which allows different types of speaking skills to be assessed. For example, Task 1 focusses on interactive skills (e.g., turn-taking), whereas Task 2 is mostly focussed on productive skills, which involve the speaker’s ability to engage in planned and rehearsed monologues (Green, 2014). Although the developers of IELTS claim that the speaking test does not allow test takers to rehearse set responses, Task 2 of the speaking module does allow them time (1-2 minutes) to plan their response. This is beneficial, as it allows not only spontaneous but also extemporaneous speaking skills to be tested, both of which are useful in the TLU domain.

Zahedi and Shamsaee (2012) have questioned the construct validity of the speaking module, which requires test takers not only to speak, but also to listen to the examiner, to read the task card, and to write notes for Task 2. However, as Douglas (2000) explains, a speaking task would be impossible without a prompt that provides explicit contextual information to the test takers. Ideally, the prompt in a speaking task should be given entirely in oral format, but the reading component in this module is not complex and/or extensive and its effects on the construct validity are negligible.

Probably the most positive feature of the speaking module is that, unlike other large-scale tests (e.g., TOEFL), it allows for face-to-face speaking to take place, which ensures both the situational and the interactional authenticity of the test. The high degree of reactivity in the speaking tasks (especially in Task 1 and 3) allows for a more naturalistic setting which prompts a more realistic performance by the assesses.

Something less positive about the speaking module concerns the evaluation criteria. In particular, foreign accent is also assessed as part of the pronunciation criterion (IELTS, 2020),

while accent does not always affect comprehensibility and intelligibility, which are the two most important aspects of L2 pronunciation skills (Derwing & Munro, 1997).

4. CHARACTERISTICS OF THE ASSESSMENT

Both the academic and the general version of the test use the same 9-band scale to report results. The overall band score is the average of the band scores obtained on the four different modules, and it is rounded to the nearest half or whole band. Assessment of receptive skills is straightforward: both the listening and the speaking module contain 40 questions, and each correct answer is worth 1 mark. Scores out of 40 are converted to the 9-band scale using special tables. Assessment of productive skills is done through rating scales which include the criteria described in the rubric. The descriptors of each band were developed by experienced raters who identified sample scripts for each level and key features of each script (Council of Europe et al., 2001). Since the IELTS test was developed before the Common European Framework of References (CEFR) and it uses a broader proficiency continuum, there is not an exact correspondence between the IELTS bands and the CEFR levels. Figure 2 is an approximate representation of how the IELTS scores align with the CEFR levels (IELTS, 2020).

CEFR	IELTS	
C2	9	PROFICIENT USER
	8	
C1	7.5	
	7	INDEPENDENT USER
B2	6.5	
	6	
B1	5.5	BASIC USER
	5	
A2	4.5	
A1	4	

Figure 2. *Approximate relation between IELTS and CEFR levels*

There is not a pass or fail result on the IELTS test, which allows institutions to set their own standards and minimum scores. A band score of 6.5 is a typical cut-off score for many undergraduate programs (University of Calgary, 2020).

Regarding the scoring criteria, it is not clear why the Academic Reading test has lower cut-off scores (i.e., it requires a lower number of correct answers) for each band compared to the General Training Reading test. It is plausible that the academic version includes more complex texts, and test takers tend to score lower. However, it serves a different purpose, and scores on the two versions do not need to be comparable. Finally, while the regular training IELTS examiners are provided with is supposed to increase the reliability of the test scores, the results would be even more reliable if productive tasks (i.e., those tasks that allow for a certain degree of subjectivity) were marked by more than one test scorer (Uysal, 2010).

5. CONCLUSION

IELTS has played a key role in assessing English language proficiency for almost 30 years. Studies suggest that the results from the IELTS test are significantly correlated with those of other major language proficiency tests (e.g., TOEFL), which demonstrates the criterion-related validity of IELTS (e.g., ETS, 2010; Zahedkazemi, 2015). Moreover, the IELTS test is fast to administer (i.e., completed in less than 3 hours), and since 2016 it is available in selected countries in both paper and computer-based format (IELTS, 2020) – a feature that undoubtedly improves its practicality and contributes to its increasing popularity. While the listening component might have a lower level of situational authenticity, the face-to-face assessment in the speaking module is a positive characteristic that is retained even for test takers who opt for the computer-based version of the test. The standardized test-taking conditions and assessment criteria of a large-scale test like IELTS ensure the reliability of the test. Assessment that works well in one context, however, may not work equally well in another, and test-takers of different learning backgrounds may be used to different accents and different types of assessment (Green, 2014). Nevertheless, IELTS is an international test and it includes a variety of accents and task types, which ensures the consistency of the results.

The present review demonstrated that there are many ways in which the IELTS test can be improved (e.g., increasing the authenticity of the listening modules and reducing the role of construct irrelevant skills). While it is far from flawless and not the only option, IELTS continues to be one of the most popular international tests of English language proficiency. Clearly, the test is an important gate-keeping measure and an incentive for millions of non-native speakers to improve their English language skills. As we know, the beneficial consequences of a given assessment system are on the top of the hierarchy of effective assessment characteristics (Green, 2014), and IELTS seems to achieve its purpose. However, it is hoped that the present critical review is a valuable contribution to the ongoing validation and improvement of the test. At the very least, it is hoped that it would help assessment stakeholders to better understand the structure of the test and to reflect on its usefulness in a more informed and objective way.

REFERENCES

- Alavi, S. M., & Masjedlou, A. P. (2017). Construct under-representation and construct irrelevant variances on IELTS academic writing task 1: Is there any threat to validity? *Theory and Practice in Language Studies*, 7(11), 1097-1107. <http://dx.doi.org/10.17507/tpls.0711.19>
- Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the international English language testing system (IELTS) listening module. *International Journal of Listening*, 26(1), 40-60. <http://dx.doi.org/10.1080/10904018.2012.639649>
- Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

- Cooper, T. (2013). Can IELTS writing scores predict university performance? Comparing the use of lexical bundles in IELTS writing tests and first-year academic writing. *Stellenbosch Papers in Linguistics Plus*, 42, 63-79. <http://dx.doi.org/10.5842/42-0-155>
- Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division (Strasbourg). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.
- Derwing, T.M. & Munro, M.J., (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1-16. <https://doi.org/10.1017/S0272263197001010>
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- ETS (2010). *Linking TOEFL iBT scores to IELTS scores – A research report*. Princeton, NJ: Educational Testing Service. Retrieved June 9, 2020 from: https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf.
- Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test. *Assessing Writing*, 10(1), 44-60. <http://dx.doi.org/10.1016/j.asw.2005.02.002>
- Green, A. (2007a). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice*, 14(1), 75-97. <http://dx.doi.org/10.1080/09695940701272880>
- Green, A. (2007b). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge: Cambridge University Press.
- Green, A. (2014). *Exploring language assessment and testing: Language in action*. Abingdon, Oxon: Routledge.
- Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia - Social and Behavioral Sciences*, 98, 637-644. <http://dx.doi.org/10.1016/j.sbspro.2014.03.462>
- IELTS. (2020). *IELTS Home of the English Language Test*. [online] Available at: <https://www.ielts.org/> [Accessed 29 November 2020].
- Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4(1), 43-66. <http://dx.doi.org/10.1016/j.jeap.2004.02.001>
- Pawlikowska-Smith, G., Centre for Canadian Language Benchmarks, & Canadian Electronic Library (Firm). (2013). *Theoretical framework for the Canadian language benchmarks and niveaux de compétence linguistique canadiens*. S.l.: Centre for Canadian Language Benchmarks.
- Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal*, 64(3), 314-320. <http://dx.doi.org/10.1093/elt/ccp026>
- University of Calgary. (2020). *A.11 English Language Proficiency*. [online] Available at: <http://www.ucalgary.ca/pubs/calendar/current/a-11.html> [Accessed 29 November 2020].
- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, 90(1), 6-18. <http://dx.doi.org/10.1111/j.1540-4781.2006.00381.x>

- Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability*, 24(3), 263-277. <http://dx.doi.org/10.1007/s11092-011-9137-z>
- Zahedkazemi, E. (2015). Construct validation of TOEFL-iBT (as a conventional test) and IELTS among Iranian EFL test-takers' performance on speaking modules. *Theory and Practice in Language Studies*, 5(7), 1513-1519. <http://dx.doi.org/10.17507/tpls.0507.27>