An evaluation of the IELTS Academic Writing subtest: Practicality, reliability & validity

By Errol Pitts, Canada

Jointly owned by the British Council, IDP ILETS Australia and Cambridge ESOL, the International English Language Testing System, more commonly known as the IELTS test, is among the most widely recognized English language proficiency tests in the world. Indeed, it is the most popular visa and citizenship test in Australia, Canada, New Zealand and the United Kingdom (British Council, 2019); and it is accepted by all the universities in Australia and the United Kingdom, more than 3,400 post-secondary institutions in the United States and hundreds of others around the world (British Council, 2019). Serving governments and post-secondary institutions, the IELTS test has, accordingly, two versions, or modules: the General Module, which governments use for visa and immigration purposes; and the Academic Module, which post-secondary institutions use to verify the English language proficiency of international students for undergraduate and graduate admissions. Though serving different purposes, both modules have four subtests, one subtest for each language skill (i.e., listening, speaking, reading and writing). Although each subtest carries equal weight in the overall calculation of a total IELTS test score, for post-secondary institutions, the writing subtest of the Academic Module (AWST) is of particular salience because, as Uysal (2010) states, writing is central to success in university. Many university assignments are, after all, written compositions, discussion board postings, reports, and essays, for example. Thus, the AWST stands out as a high stakes test, playing a key role in the determination of who can and who cannot pursue their personal and professional aspirations in academia. Thus, the AWST merits an evaluation to determine to what degree it serves as an effective language assessment.

Though debate about how to evaluate language assessments exists, scholars have identified five to six qualities of an effective language assessment. For example, Bachman and Palmer (1996) propose that test usefulness requires a balance between six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Among these qualities though, Bachman and Palmer emphasize the importance of reliability and validity. Along a similar line, Brown (2004; as cited in Thuy & Dung, 2020) emphasizes five criteria of an effective language test: practicality, reliability, validity, authenticity, and washback. Like Brown, Hughes et al. (2020) also highlight five qualities: practicality, reliability, validity, impact, and backwash. Though differing slightly, these three approaches to evaluating the effectiveness of a language assessment all share the aspects of practicality, reliability, and validity. Given this consensus among scholars, I therefore use these three characteristics to evaluate the AWST.

Surprisingly, despite its widespread recognition by post-secondary institutions around the world and its significant attention in the admissions process, there is a death of literature investigating the effectiveness of the AWST. To contribute to this shortage, I evaluate it. More specifically, in this evaluation, I provide an overview of the AWST, and then examine it drawing on literature and research with respect to practicality, reliability, and validity.

Overview of the IELTS Academic Writing Subtest

The AWST consists of two distinct tasks: Task 1 and Task 2, both of which must be completed in under 60 minutes. Task 1 is a report of at least 150 words and requires a description, explanation, or summary of a visual graphic: a graph, chart, table, diagram, map, or process chart. The report must have an introduction and overview and highlight and compare the main features or data (IDP IELTS, 2023a). To illustrate with an example, a test candidate may need to analyze a line graph that visually compares the number of men and women studying online and face to face in Canadian higher education across three time periods.

Unlike Task 1, Task 2 is an argumentative essay of at least 250 words and requires a response to a point of view, an argument, or a problem focused on a relevant topic. The essay must have an introduction, body, and conclusion, and "it is important that [the test candidate] complete[s] the task carefully using relevant ideas and examples to support [their] position. [Their] ideas should be organised clearly, using paragraphs for each idea" (IDP IELTS, 2023b, para. 3). For an example of Task 2, a test candidate may need to argue for or against a proposed law that prohibits homeschooling.

Evaluation of the IELTS Academic Writing Subtest

Practicality

The practicality of a test can be defined as the extent to which it is economical, easy to make, easy to score, and easy to interpret (Cervatiuc, 2023). Regarding the economics of the AWST, literature that describes the time and costs involved with its design and development is unavailable or difficult to locate on the internet; however, literature that summarizes its development and validation processes (IELTS, 2023, para. 1) is available, though rather general in nature. To explain, drawing on this general information from the IELTS (2023) webpage, the test design and development process can be construed to consist of the following five stages. First, IELTS commissions teams of language specialists based in Australia, Canada, New Zealand, the UK, and the USA to write prompts that adhere to test specifications. Second, IELTS pre-edits the commissioned prompts, and, if necessary, provides suggestions to the language specialists to revise the prompts. Third, IELTS reviews the edited prompts and either approves them for pre-testing or provides additional edits. Fourth, the IELTS Validation Team pretests the prompts on representative samples of test candidates to determine whether the prompts can differentiate between strong and weak test candidates. Fifth, IELTS constructs the tests considering multiple factors, such as, providing a range of cultural perspectives, and balancing task types, topics, and genres. To sum up, involving multiple stakeholders around the world engaging in a series of tasks, the test design and development process is potentially challenging and likely to require much time and money.

How easy is the AWST to score? Research shows that IELTS examiners have mixed perceptions. For example, a Cambridge ESOL study into the writing subtests of both the Academic and General Modules, which was conducted by Shaw and Falvey (2008), found that nearly all the IELTS examiners in their study believed that the scales are clearly worded, and nearly three-quarters of the examiners stated that the scales are easily interpretable. In a contrasting study, however, Mickan (2003) found that examiners struggled to find lexico-grammatical features that distinguish different levels of performance using the IELTS band descriptors. Mickan's study did focus on the writing test of the General Module though, not the Academic Module. Nevertheless, considering that in the General and Academic writing subtests, Task 2 is an essay, and the Task 2 rubrics are very similar for both Modules, the findings of Mickan's study does warrant consideration. Along the same line, some of the descriptors in the AWST Task 1 and Task 2 rubrics are somewhat vague, which can result in difficulty distinguishing different levels of performance. Consider, as evidence, the following band 8 and 9 descriptors under lexical resources on the Task 2 rubric: "Occasional errors in spelling and/or word formation may occur, but have minimal impact on communication" (IELTSTutors, 2023, p. 7), and "minor errors in spelling and word formation are extremely rare and have minimal impact on communication" (IELTSTutors, 2023, p. 7). Differentiating between *occasional errors may occur* and *minor errors are extremely rare* can indeed be challenging.

Reliability

A characteristic of test data, reliability can be described as consistency of test data across repetitions of test administration (Chapelle, 2013). Put another way, if a test yields similar results in several administrations with the same students, the test data is reliable.

To foster reliability, IELTS implements several appropriate training and certification processes. For instance, IELTS subjects writing prompts to trials on sample populations of test candidates, which, according to Chapelle (2013), builds reliability. Additionally, IELTS provides rigorous initial training and ongoing certification processes for its examiners. More specifically, potential examiners undergo standardized training by an experienced IELTS examiner trainer, and this training requires potential examiners to score IELTS writing tests for 1.5 days. This training also includes norming or standardization training, which, according to Weigle (1994), has been shown to help examiners apply rubrics in their intended ways. Furthermore, norming, according to Jacobs et al. (1981; cited in Weigle, 1994, p. 198), has also been found effective in neutralizing the effects of the backgrounds of examiners, a particularly beneficial effect since examiners' linguistic and cultural backgrounds are diverse.

With respect to the quality of IELTS training, two studies evaluate it highly. First, McDowell (2000) investigated the perceptions of IELTS examiners about IELTS training and found that most examiners are satisfied with the training, but some did state that they want more time to 'digest' the band descriptors, giving weight to Mickan's (2003) study regarding the perception that some examiners found the descriptors difficult to distinguish. Second, in an IELTS Australia and British Council published report, Cotton and Wilson (2008) investigated cohesion and coherence in Task 2 and conclude that

> No effect could be found for IELTS marking experience, higher qualifications, training in linguistics, and either the level of most teaching experience or for the number of years of teaching experience. This would seem to suggest that the IELTS training, certification and re-certification processes have been effective in ensuring the reliability of examiners regardless of differences in their background and experience. (p. 49)

In addition to trialing prompts on sample populations and providing high quality training, potential IELTS "…examiners must mark a series of exams consistently and accurately" (IDP IELTS, 2023b, para. 2) to become certified examiners, and all examiners are re-certified every two years to ensure standards are being maintained (IDP IELTS, 2023b). Furthermore, IELTS requires examiners to use analytic scales to score writing tests. An analytic scale "…includes a number of separate criteria to guide the assessor's judgements and so generates a number of different scores – e.g., a score for grammar, a score for pronunciation and a score for task fulfilment" (Green, 2020, p. 252). Analytic scales, according to Shaw and Falvey (2008), provide advantages over global scales as they provide enhanced reliability through increased observations, encourage greater discrimination across a wider range of assessment bands, and discourage norm-referencing.

Despite these strategies to build reliability in the AWST, Uysal (2010) challenges the AWST's inter-rater reliability, that is, the extent to which the same test re-scored by a different rater yields similar results. Uysal challenges the inter-rater reliability because the AWST are single scored locally, and double scoring, according to Green (2014), is one way to build reliability. On this point however, IELTS does subject selected samples of scored writing subtests to be double scored (Uysal, 2010). That is, after a test candidate has completed their test at a local IELTS centre, IELTS selects samples from these test centres, and then senior examiners double score the samples. Despite the fact that most tests are still single-scored though, the correlation between local and senior raters was 0.91 in 2003 (Hashemi & Daneshfar, 2018; Veerappan & Sulaiman, 2012), which indicates high reliability (Chapelle, 2013).

Validity

Although it produces reliable test data, the AWST falls short of demonstrating a high degree of validity, a complex quality which is, according to Messick (1989; cited in Fulcher & Davidson, 2007), the quality of a test that provides a teacher with a degree of justification to make inferences from a test result to a test construct. Three kinds of validity are briefly examined: content, construct, and predictive validity.

Content validity refers to how well a test "…constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned" (Hughes, 2003, p. 26), and the AWST is concerned with, but is somewhat unrepresentative of, post-secondary writing assignments. To this point, several studies serve as evidence. For instance, Moore and Morton (2005) compared the Task 2 rubric to writing assignments in two courses in two Australian universities and found that, although there are some similarities between Task 2 and university assignments, there are "…important differences between the writing required at university and that required to pass the IELTS test" (p. 63). In a similar vein, Cooper (2014) compared lexical bundles, or groups of words that occur frequently, between first-year undergraduate essays and samples of academic Task 2 writings, and found that

> there are notable differences in the structures and functions of lexical bundles used in Task 2 of the IELTS writing test and in students' undergraduate assignments. The nature of these differences suggests that the style of writing expected in the IELTS test does not correspond to that expected in student assignments. (p. 78)

Another differentiating characteristic between the AWST and university writing tasks is that university writing assignments typically require students to draw on external resources (e.g., academic studies, reports and books) while Task 2 requires test candidates to draw on personal knowledge and experience (Nguyen, 2016), a flaw that also connects to construct validity.

Construct validity, according to Cervatiuc (2023), is the ability of a test to measure what it intends to measure, and, as its name implies, the AWST intends to measure academic English writing proficiency. To that extent, it is rather effective, but, like all assessments, it is imperfect. Namely, there is an issue with Task 2: Task 2 is designed to measure academic writing skills such as the ability to argue and support an opinion, but Task 2 also evaluates non-academic writing skills. To highlight these non-academic skills, some Task 2 prompts instruct test candidates to use relevant examples from their own knowledge or experience. Thus, test candidates' knowledge and experience play a role in the result (Nguyen, 2016), which is problematic because "in language testing, we are not normally interested in knowing whether students are creative, imaginative or even intelligent, have wide general knowledge, or have good reasons for the opinions they happen to hold. For that reason, we should not set tasks which measure those abilities" (Hughes, 2003, p. 82; cited in Weigle, 2002, p. 45). Therefore, the construct validity requires improvement.

Although the AWST's content and construct validity require attention, research studies about its predictive validity, "…the term used when the test scores are used to predict some future criterion, such as academic success" (Fulcher & Davidson, 2007, p. 5), are mixed. For example, an ESOL Cambridge study by Kerstjens and Nery (2000) investigated the ability of the IELTS test to predict the academic success of first-year international students in an Australian university and found that there was a weak positive correlation between the reading and writing tests and GPA. Another study by Yen and Kuzma (2009) investigated the predictive validity of the AWST and found that there is a significant correlation between first-year students' GPAs and their AWST scores. In a different vein though, Mauriyat (2021) reviews literature about the AWST's predictive validity and concludes that the AWST demonstrates low predictive validity.

Conclusion

In conclusion, the AWST is a globally recognized high-stakes English language test with strengths and shortcomings. Although scholars evaluate language assessments differently, three qualities that are acknowledged by some scholars are practicality, reliability, and validity. With respect to practicality, the test development and validation processes of the AWST are arguable time consuming and expensive; however, considering the widespread recognition and administration of the IELTS test, potentially profitable. The ease of scorability of the AWST is, however, doubted. One possible reason for this doubt lays in the rubrics' descriptors, some of which may be difficult to differentiate. To address that potential shortcoming, IELTS should investigate further IELTS examiners' perceptions of the descriptors to determine whether descriptors need rewording.

About reliability, AWST data has been found to be reliable. Potential reasons for this strength include the trialing of prompts on sample populations of test candidates, high quality training and continuous certification processes, and the use of analytic scales. Although the AWST produces reliable data, research presents issues with its content and construct validity. Specifically, research highlights that the AWST fails in some ways to represent university writing assignments, tainting the content validity. Research also points out that non-academic writing skills and knowledge are assessed in the AWST, polluting the construct validity. Taken together, that is, the issues with its content and construct validity, post-secondary institutions should interpret results of the AWST cautiously. In fact, such institutions should consider designing, developing, and administering an internally and locally constructed writing test to measure academic writing skills. With respect to predictive validity, research indicates slightly more positive than negative results. Regardless, when viewing the results of the AWST through the lens of predictive validity, post-secondary institutions need to be vigilant because a host of issues, in addition to language proficiency, affect the future academic success of international students, for instance, financial matters, cultural differences, and personal traits. Though research has presented issues with the validity of the AWST, as Uysal (2010) mentions, "IELTS is committed to improving the test further [IELTS test] and has been carrying out continuous research to test its reliability and validity" (Uysal, 2010, p. 319).

References

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

British Council. (2019, March 19). IELTS. *IELTS grows to 3.5 million a year*. https://takeielts.britishcouncil.org/about/press/ielts-grows-three-half-million-year#:~:text=IELTS%20is%20the%20International%20English,taken%20in%20the%20last%20year

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.

Cervatiuc, A. (2023). *The characteristics of effective language assessments* [PowerPoint]. https://canvas.ubc.ca/courses/135083/pages/content-of-module-2?module_item_id=5943046

Chapelle, C. A. (2013). Reliability in Language Assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–6). https://doi.org/10.1002/9781405198431.wbeal1003

Cooper, T. (2014). 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*(1), 63–79. https://doi.org/10.5842/42-0-155

Cotton, F., & Wilson, K. (2008). *An investigation of examiner rating of coherence and cohesion in the IELTS Academic Writing Task 2* (IELTS Research Report 12; pp. 1–76). IELTS. https://ielts.org/researchers/our-research/research-reports/an-investigation-of-examiner-rating-of-coherence-and-cohesion-in-the-ielts-academic-writing-task-2

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment* (1st ed., Vol. 4). Routledge. https://www.kau.edu.sa/Files/0005056/Subjects/Fulcher%20Davidso%20Language%20Testing%20and%20Assessment%20An%20Advanced%20Resource%20Book%20Routledge%20Applied%20Lingu.pdf

Green, A. (2014). *The qualities of effective assessment* (Second). Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315889627/exploring-language-assessment-testing-anthony-green

Green, A. (2020). *Exploring language assessment and testing: Language in action* (Second). Routledge. https://www.routledge.com/Exploring-Language-Assessment-and-Testing-Language-in-Action/Green/p/book/9781138388789?srsltid=AfmBOoqMfow_BDw7ZNPdtcg71Uyj8HzquT_xirsBzf3Eg2YF3ujAgyFo

Hashemi, A., & Daneshfar, S. (2018). A Review of the IELTS test: Focus on validity, reliability, and washback. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *3*(1), 39–52. https://ijeltal.org/index.php/ijeltal/article/view/123

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511732980

Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (Third). Cambridge University Press. https://gw2jh3xr2c.search.serialssolutions.com/?sid=sersol&SS_jc=TC0002682903&title=Testing%20for%20language%20teachers

IDP IELTS. (2023a). *IELTS Academic Writing test practice questions*. IELTS. https://ielts.idp.com/prepare/article-academic-writing-free-practice-questions

IDP IELTS. (2023b). *Who are IELTS examiners and how do they score the Tests?* IELTS. https://ielts.idp.com/canada/prepare/article-understanding-ielts-examiners-and-how-they-score-tests

IELTS. (2023). How we develop the test. IELTS. https://ielts.org/organisations/ielts-for-organisations/how-we-develop-the-test#:~:text=Commissioning,components%2C%20and%20outline%20specific%20requirements.

IELTSTutors. (2023). *IELTS public writing band descriptors*. IELTS Tutors. https://ieltstutors.org/writing-band-descriptors/#:~:text=In%20the%20writing%20test%2C%20an,1%20and%20task%202%20test

Kerstjens, M., & Nery, C. (2000). *Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance* (Research Reports 2000 Volume 3; Research Reports 2000). IELTS Australia. https://ielts.org/researchers/our-research/research-reports/predictive-validity-in-the-ielts-test-a-study-of-the-relationship-between-ielts-scores-and-students-subsequent-academic-performance

Mauriyat, A. (2021). Authenticity and validity of the IELTS writing test as predictor of academic performance. *PROJECT (Professional Journal of English Education)*, *4*(1), 105. https://doi.org/10.22460/project.v4i1.p105-115

McDowell, C. (2000). *Monitoring IELTS Examiner Training Effectiveness: A preliminary study* (ILETS Research Reports 2000 Volume 3; pp. 1–141). IELTS Australia. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/monitoring-ielts-examiner-training-effectiveness-mcdowell-2000.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 12–103). Macmillan/American Council on Education.

Mickan, P. (2003). *What's your score? An investigation into language descriptors for rating written performance* (Volume 5, Paper 3). IDP IELTS Australia. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/whats-your-score-mickan-2003.pdf

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. https://doi.org/10.1016/j.jeap.2004.02.001

Nguyen, H. T. M. (2016). A review of the IELTS writing test. *Journal of Science and Technology - The University of Danang*, *10*(107), 22–37. https://jst-ud.vn/jst-ud/article/download/1794/1794/6221

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Towards a revised rating scale*. University of Cambridge ESOL Examinations. https://www.researchgate.net/publication/267362396_The_IELTS_Writing_Assessment_Revision_Project_Towards_a_revised_rating_scale

Thuy, N. N., & Dung, N. L. T. (2020). Major language test qualities and ways of enhancing them. *Research Journal of English Language and Literature*, *8*(2), 1–7. http://www.rjelal.com/8.2.20/1-7%20NGUYEN%20LUONG%20TUAN%20DUNG.pdf

Back

Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal*, *64*(3), 314–320. https://doi.org/10.1093/elt/ccp026

Veerappan, V., & Sulaiman, T. (2012). *A Review on IELTS Writing test, its test results and inter rater reliability*. Theory and Practice in Language Studies, *2*(1), 138–143. https://doi.org/10.4304/tpls.2.1.138-143

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Yen, D., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching*, *3*, 1–7. https://rteworcester.wp.worc.ac.uk/wp-content/uploads/2017/05/yenkuzmaieltscores.pdf

Author Bio

Errol Pitts has been teaching for over 20 years. He has an MEd and a BEd from the University of Manitoba, and a BSc from the University of Winnipeg.

Back