Bulletin of Advanced English Studies (BAES)

Volume.8, Issue.2, pp: 72-92
Dec (2023)

Article 4

Examining the Reliability of the International English Language Testing System Design

Michael Suss

The Imperial College of Australia, Australia
ceo@imperal.edu.au

Received: 30/5/2023 Revised: 24/6/2023 Accepted: 12/7/2023

DOI: https://doi.org/10.31559/BAES2023.8.2.4

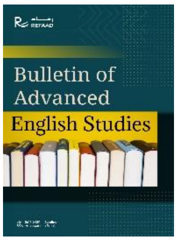
REFAAD

www.refaad.com

Bulletin of Advanced English Studies (BAES)

Journal Homepage: https://www.refaad.com/Journal/Index/5

E-ISSN 2617-6459 | P-ISSN 2617-6440



Bulletin of Advanced English Studies

Examining the Reliability of the International English Language Testing System Design

Abstract:

Background: Despite the evolution of the International English Language Testing System (IELTS) to accommodate the evolving needs of test-takers and institutions, concerns have been raised regarding its reliability in assessing English language proficiency.

Purpose: This study aimed to investigate the reliability of the International English Language Testing System (IELTS) design as a measure of English language proficiency.

Methodology: To explore this issue, the research reviewed relevant literature and employed various methods, including frequency analyses, data analytics, and machine learning algorithms based on mathematical principles. Unlike previous studies, this work analysed a dataset of 33,505 authentic band scores, allowing for a comprehensive examination of IELTS test score patterns.

Findings: The findings revealed that the concept of reliability is complex, and the statistical methods used by IELTS test providers, based on classical test theory, may provide approximate rather than precise scores. Relying solely on the reputation and authority of the IELTS test is insufficient to establish its reliability, as it involves logical fallacies and arguments from authority.

Research implications: It is necessary to conduct further independent research to ensure the IELTS test's reliability and address the concerns expressed by researchers and test-takers.

Keywords: *IELTS; reliability; test; Australia; assessment.*

1 Introduction

The IELTS test, also known as the International English Language Testing System, Is used to measure the English language proficiency skills of English language test candidates. It is commonly used by many organizations globally to evaluate an individual's English language proficiency skills.

The IELTS exam's Academic module is targeted toward individuals who plan to attend universities or tertiary institutions where English is the language of instruction, making it a valuable tool for those interested in pursuing higher education (British Council, 2023) in countries such as Canada, Australia, the U.K., or New Zealand. This module evaluates the candidate's capability to comprehend academic texts, compose academic essays, and participate in academic discussions. In contrast, the General Training module caters to those planning to move to an English-speaking country for work or other reasons. This module measures the test-takers' skills in understanding everyday language, creating correspondence related to personal or professional issues, and participating in informal conversations and workplace discussions. In both modules, the test-takers' proficiency is evaluated using a 0-9 band scale, where 0 indicates no proficiency, and 9 indicates an expert level of proficiency.

In 1989, its inception year, the IELTS exam underwent several adjustments. One such modification occurred in 1995 when it was split into two modules- Academic and General Training. Both modules measure the test-takers' language ability in the four language competencies: listening, reading, writing, and speaking. Nonetheless, their purpose and target audience varied (Cambridge University Press, 2023).

Although the IELTS test is widely regarded as a good measure of language proficiency, it was not originally designed for precision, as compromise was necessary during its evolution (Clapham, 1996). This compromise balanced practicality with maximum predictive power (Alderson & Clapham, 1992). Therefore, the IELTS test's band scales may be considered approximate values from a test reliability design perspective, but they are still sufficiently accurate for their intended purpose (Everitt, 2010).

Despite the significant amount of literature on the IELTS test, it remains shrouded in secrecy, and many researchers seem content with this. One example is that the providers of the IELTS test cite commercial confidentiality as the reason for not granting independent researchers access to past test band scores and assessment tests (Cambridge Assessment, 2018). Despite this restriction, some researchers use exemplar test papers without disclosing that the IELTS test providers do not provide data from previous tests. Consequently, readers of their research are unaware that the study did not use actual IELTS test data.

1.1 Problem statement

As a registered migration agent with years of experience, I actively opposed the proposed introduction of English language testing by the Migration Institute of Australia (MIA) and the Office of the Migration Agents Registration Authority (OMARA). The requirement would have subjected registered migration agents from non-English speaking countries, as defined by the English language assessment test owners, to mandatory English language testing at the time of their registration renewal. This requirement was to occur of how long they lived in Australia, and some countries considered 'non-English speaking' had English as their lingua franca. The criterion for inclusion or exclusion seemed to be based more on race, with white Anglo-Saxon people from accepted countries and people of color, who are native English speakers, from excluded countries like Nigeria, Ghana, Kenya, Singapore, and the Philippines.

As a result of this requirement, only Registered Migration Agents who came from countries not classified as English-speaking were subject to English language testing during their registration renewal. This policy did not consider the individual's literacy skills or educational qualifications and instead relied solely on their country of origin. The Office of the Migration Agents Registration Authority (OMARA) assumed that individuals from approved countries with an English-speaking background automatically met all English language requirements. However, as an English language teacher for international students in the Vocational Education and Training (VET) sector, I disagreed with the notion that all Australians have high literacy skills. Chelliah (2010) also rejected the assumption that being a native English speaker automatically qualifies someone as competent in the English language.

To clarify, until recently, Registered Migration Agents did not have to undergo additional English language testing to renew their registration. However, if such testing were to be introduced, it is my understanding that very few agents would pass. Therefore, I perceived this proposed requirement to exclude people of color from becoming Registered Migration Agents.

The third issue with the proposed policy change was that it would have discriminated against naturalised Australian citizens who had migrated to the country several decades earlier, regardless of how long they were registered as a migration agent. Such discrimination appeared unethical in a country that upholds the principle of equality before the law.

However, this study does not focus on the legality of using the IELTS test, as that can only be determined by the Australian legal system. Rather, the study examines whether the IELTS test is an unreliable measure of English language proficiency, particularly for individuals from non-English speaking backgrounds, often people of color.

The main inquiry of this study can be formulated as follows: **Does the IELTS test provide a reliable evaluation of candidates' aptitude in using the English language?**

2 Literature Review

Many studies on the reliability of the IELTS test design that were reviewed lacked critical thinking. Instead of conducting their research, some researchers quoted and relied on the research of others, even though those studies may not have been peer-reviewed. This approach can lead to inaccurate conclusions, as the researchers may not have thoroughly examined the evidence. Researchers need to engage in critical thinking and evaluate the quality and reliability of the evidence they are using in their studies. By doing so, they can ensure that their research is based on credible and trustworthy sources, leading to more accurate and reliable conclusions.

2.1 Understanding Reliability as a Measurement's Consistency

Reliability can be measured in two ways: by a reliability coefficient (ρxx), where 0 indicates total unreliability, and 1 represents full reliability (Salkind, 2006), and by the standard error of measurement (σmeas). Using the latter makes it possible to determine a confidence level surrounding a test candidate's true score, compared to their observed score. For example, one could say that an IELTS test candidate's true score falls within an interval with a lower limit of 5.41 and an upper limit of 7.52. However, using the IELTS quantizing method, the true score would fall between Band 5.5 and Band 7.5, a confidence interval of two bands, making the IELTS test an unreliable measure of a person's English language proficiency skills. To increase the reliability of the IELTS test, one needs to reduce the confidence interval.

The IELTS test providers claimed that the IELTS test is an accurate measurement of a candidate's English language proficiency at a moment in time, meaning that the IELTS test band scores reflect the results of the test held on the test day (Cambridge ESOL, 2004; Green, 2004; IDP: IELTS Australia, 2011; IELTS, 2013, 2016). The IELTS test providers also claim that the IELTS test is fair and accurate (IELTS, 2013, 2016); IELTS is a clear and fair reflection of the test taker's ability or 'true reflection' (IELTS, 2015a, 2022); the IELTS test band scores measure the test candidate's language proficiency at a given point in time (Cambridge ESOL, 2004; Green, 2004; Hawthorne, 2013; IDP: IELTS Australia, 2011; IELTS, 2015b; Kunnan & Jang, 2009); and IELTS test band scores provide an accurate measure or picture of the candidate's English language proficiency (Cambridge ESOL, 2004; IELTS, 2009, 2015a, 2018).

Although the IELTS test providers did so, up to a few years ago, referring to the standard error of measurement (SEM), they failed to explain the implications of their test results containing an SEM. For example, Perlin et al. (2021, pp. 464-465) noted that the .S.U.S. Supreme Court, in the judgement Atkins v Florida (Hall v. Florida, 134 S. Ct. 1986, 1998 (2014)) pointed out that a test result is not a precise number. There is now universal agreement that assessment test results, for example, .Q.I.Q. tests should not be read as a single integer but to be understood to be a range. This Supreme Court judgment recognised what statisticians have always known. Similarly, the IELTS test band scores should be seen to cover a range and not as a single number. Although the IELTS test providers published the IELTS test performances annually together with the standard error of measurements (SEM), they could not be independently verified (IELTS, 2014b, 2015g, 2016l, 2019h).

Wilde (2002) emphasised the significance of test reliability as one of the fundamental aspects that arecontributing to test accuracy, with test validity being the second. However, Davies (1999) contested that test reliability pertains to the differences between the outcomes of one test itself or another, and the presence of examiner bias or variations in test conditions may account for measurement errors. On the other hand, IELTS (2004) defined reliability as the extent to which test scores are not influenced by measurement errors, highlighting that measurement errors can diminish the reliability and generalizability of the test scores obtained from a single measurement for an individual.

Test reliability definitions more acceptable for this study were those given by Carmines and Zeller (1979), Palomba and Banta (1999), and Punch (2013). They defined test reliability as the extent to which a measurement gives consistent results, with higher consistency indicating higher test reliability. Fitzner (2007) emphasised that high test reliability meant consistently similar test results. Test reliability is measured on a scale of 0 to 1, with 1 indicating no error or variation. Mokhsin et al. (2015) defined test reliability as the degree of consistency that the test instrument measures the variable to be measured in the research study.

Sawand et al. (2015) highlighted two principles that must be recognised concerning test reliability. Firstly, it pertains to test scores, not the test instruments, such as the IELTS test analysed in this study. Secondly, the error size tends to be underestimated, and thus the reliability of the IELTS test may not be entirely accurate. Consequently, other statistical methods may also be unsuitable for this study, including the standard error of measurement (SEM), reliability coefficients, and confidence intervals, which involve approximations and rounding of numbers. However, it is crucial to note that several researchers, such as Pearson (2019a), Templer (2004), and Uysal (2009, 2010), have raised concerns about the IELTS test or exemplar test assessments, although some of these concerns have only been addressed by IELTS-aligned researchers (Green, 2019; Hall, 2009). It seems to be an unwritten rule that any critical analysis will not be tolerated, and researchers are expected to accept published and cited information without question, as Watson and Hayter (2020, 3) argued that "if it's published, it must be true." However, one study should not form the basis for conclusions.

Furthermore, according to Davies (2005), universities and other institutions adopted the ELTS and the IELTS tests because they believed in their reliability. However, he also noted that even if the tests were not entirely reliable, additional costs would still be associated with their use. Davies (2005) explained that since no test can have complete reliability, it was necessary to estimate, as accurately as possible in financial terms, the impact of using an imperfect instrument on the performance of overseas university students. This dilemma is similar to the questions addressed in medical reports that consider the acceptable proportion of false positives and negatives resulting from using a new drug.

From the preceding, the IELTS test providers claim that its reliability is based on test data from 2009, as stated in their publications (IELTS, 2014a, 2015a, 2016 a). However, whether this refers to data from every year after 2009 or just from that year alone is unclear. Additionally, these claims made by the IELTS test have not been independently verified, nor can they be replicated. Yet, in evidence-based research, researchers are generally expected to provide enough information to allow their study to be replicated. This helps to ensure research integrity and reduce measurement errors (Peng, 2011; Poldrack et al., 2017; Resnik & Shamoo, 2017). However, Resnik (2002) expressed concern about the growing amount of peer-reviewed research that was not reproducible, while Poldrack et al. (2017) were troubled by the lack of research replication and the superficiality of research quality. Peng (2011) asserted that research reproducibility was the minimum requirement for a scientific claim to be made.

Thus, this section examined the reliability of the IELTS test and noted a lack of empirical studies on its reliability using scaled IELTS test band scores. The IELTS test providers claim that their research library supports the test's reliability, but there has been no comprehensive critical evaluation of the original IELTS test data in those research reports. The section also discussed the discrepancy between the previous and current descriptions used by the IELTS test providers and mentioned the challenges of accessing actual test data for research purposes.

2.2 Confirming the Discrepancies of IELTS' Reliability

Gagen (2019), in his Master's study, claimed to have conducted a meta-analysis of 3,265 research papers that investigated the IELTS test. However, further investigation revealed that only the abstracts of these papers were indexed for analysis. The study included only nine IELTS Research Reports out of over 120 residing in the IELTS research library. The included reports were authored by Arrigoni and Clark (2015), Breeze and Miller (2011), Coleman et al. (2003), Cotton and Conrow (1998), Hill et al. (1999), Humphreys et al. (2012), Ingram and Bayliss (2007), Kerstjens and Nery (2000), and Lloyd-Jones et al. (2012).

Moreover, according to Everitt (2010), the IELTS test providers believed that their tests were sufficiently accurate for their intended purpose regarding test reliability. However, Douglas (1990) argued that admissions staff needed to exercise some flexibility in determining cut-off scores for assessment tests such as the Test of English as a Foreign Language (TOEFL), as these tests were too imprecise to support accurate discrimination and ultimately, decisions had to be made. Davies (1984) stated that the predecessor to the IELTS test, the ELTS test, was designed to possess high predictive power as a starting point for a language proficiency test. Additionally, Davies (1984) highlighted that international students were more concerned about their academic success in their field of study and having sufficient English proficiency rather than possessing near-native control over the English language. Davies (1967) had made this observation almost 20 years earlier.

Despite referring to the IELTS test as "fair, accurate, and relevant" (IELTS, 2009, 4) in the past, the IELTS test providers later changed the description to "fair, reliable, and valid," (IELTS, 2011; 2015b, 2; 2019a, 4; 2020a) without providing any evidence-based research to support this claim. The literature suggested that this new terminology does not align with the typically used concepts of test reliability, validity, and fairness in applied linguistics.

Similar to other studies, Ellis et al. (2013) encountered the same restrictions in accessing actual test data when examining the reliability of the IELTS test. As a solution, Cambridge Assessment (2018) suggested that researchers utilise exemplar test data and questions, along with the IELTS Research Reports housed in the IELTS research library.

Consequently, the IELTS test providers clarified their stance to safeguard their commercial position (Hogan, 2005; British Council et al., 2006; Cambridge Assessment, 2018). In a circular letter to all migration agents, IDP: IELTS Australia referred to their test as being "fit for purpose," presumably about the reliability and validity of the IELTS test, and supported by their research library (IELTS, 2021a). The IELTS owners, including IELTS Australia, the British Council, and the University of Cambridge ESOL Examinations, regularly invest in research to enhance the test's quality and ensure that it remains fit for purpose and support ongoing system and security enhancements. Bachman and Palmer (1996) expanded on this and incorporated further concepts under "usefulness for a particular testing context" or test usefulness.

Harvill (1991) cautioned that researchers often conflate the concepts of test reliability, and while the reliability coefficient may offer an idea of measurement errors in a group, it cannot be used to interpret individual test scores. For individual scores, Harvill (1991) recommended using the SEM. O'Loughlin (2013) defined jagged profiles as test candidates' results varying by two or more band scores. The IELTS (2004) claimed that their computerised processing package (EFLCOMMS) was programmed to identify any untypical profiles, including jagged profiles, in the four IELTS macro skills. The IELTS test providers maintained that this adjustment of test scores further demonstrated the IELTS test's reliability (IELTS, 2012, 2014 b).

What is more, Hamid (2015) conducted a study on IELTS test retakers, and while questioning the IELTS rounding-up procedures, no one raised concerns regarding the integrity of the IELTS test. He believed that although the margin may create the impression that IELTS scores are highly reliable, there are other ways to assess the value of fractions and their underlying reliability. One such example is the rounding up of scores, which can result in fractions losing their significance that benefited or disadvantaged test-takers. For example, on August 13th, Ieltser received an overall band score of 7 despite being 0.5 points short in speaking, while on October 29th, she received a score of 7 even though she had a surplus of 0.5 points in one skill. This example showed how fractions can be both significant and insignificant when it comes to IELTS scoring.

According to Fulcher and Davidson (2009), the IELTS test providers did not clarify whether the changes made to the score reporting system had any impact on the reliability of the test. However, the IELTS (2007a) addressed their query on why score reporting for the writing and speaking tests was changed, stating that the change was introduced based on feedback from many organizations and users of IELTS scores. The IELTS claimed that the half-band scores were found to help specify required language levels more accurately and that many teachers and test-takers expressed a desire for more detailed information on performance in each skill. It is worth noting that the IELTS did not offer any evidence to support these claims.

Therefore, this section highlighted the difference between the previous and current descriptions used by the IELTS test providers, which do not align with the typical concepts of test reliability in applied linguistics. It was also noted that the IELTS test providers do not disclose the mechanics behind the score and how they can impact the reliability of the test.

3 Methodology

This research follows the Mertens transformative-emancipatory paradigm, which emphasises conducting research that is non-discriminatory, participatory, and focused on marginalised individuals and groups (Johnson, 2007). Instead of solely concentrating on the IELTS test and its users, this study recognized that many people came to Australia seeking refuge from various forms of discrimination. Mertens (2003) highlights the ethical obligation of researchers to pursue new knowledge and use it to promote social change, which can empower less powerful individuals to advocate for their needs and contribute to societal transformation.

3.1 Instruments of the Study

The study's data collection was limited to January 1998 and 2013, and the selection was based on availability. The sample consisted of test candidates of different nationalities, genders, ages, academic achievements, and reasons for taking the IELTS test. However, there were limited publicly available IELTS test results after 2013. Nonetheless, the researcher believed having “some data was better than having none at all” (Persaud & Dagher, 2021,11). Yet, this study is not concerned with determining the correct standard error of measurement for the different test samples but rather highlights that the reliability of the IELTS test cannot be precisely measured. Different researchers have proposed varying levels of acceptable measurement errors for high and low-stakes tests.

3.2 Procedure: Machine Learning Program and Orange Data Mining

This study utilised SPSS and Microsoft Excel to analyse and synthesise the database, along with various techniques such as frequency analysis tables, logistic regression, and artificial intelligence ML algorithms. Numerical patterns were observed during an initial examination of the database, which may have been due to the randomised collection of test results or an unknown IELTS algorithm. The algorithm's function was explored through ML cross-validation.

The researcher employed logistic regression analysis and ML algorithms to reverse-engineer the IELTS test band scores and create a learning system. The process involved inputting data on a defined target, with the flow chart in Fig. 1. Four models were built with a single input variable, "SWRL," representing the four IELTS macro skills. The models calculated "correct predictions" to measure the contributions of each macro skill and whether they were equally weighted. Fig 1 provides an example of one of the many Orange data workflow models utilised during the data mining of the database.

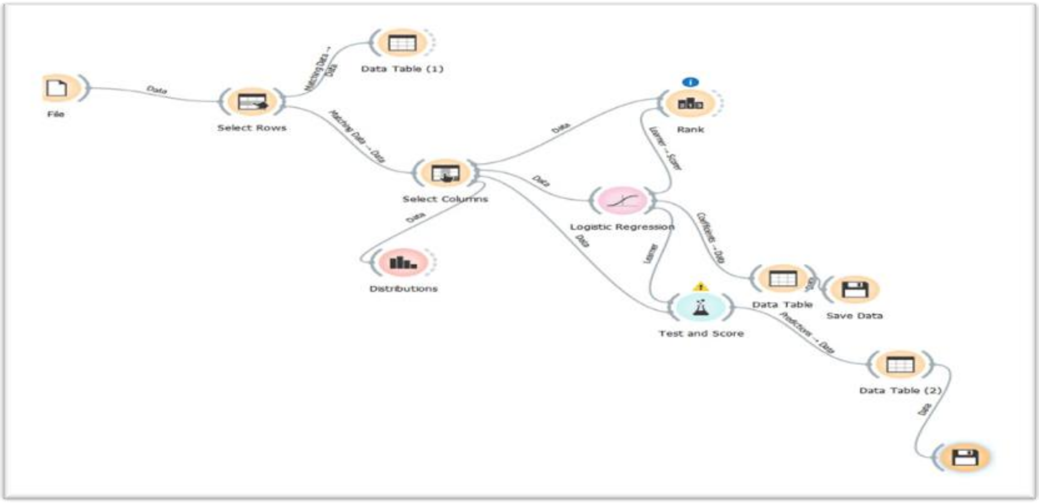


Fig. 1: The Orange Data Workflow Model Used in the Data Mining of the Research Study

In this study, ML learning was used to build intelligent models, such as a predictor for the overall band score, and to uncover patterns and relationships within the dataset. This method is particularly useful in allowing the researcher to identify which factors influenced the target scores at specific data points and gain insight into the decision-making process.

To complete the reverse-engineering of the database's frequency analysis results, Kling et al. (2007) recommended introducing Python programming. As a result, a Python programmer was hired to make the pickle file readable and develop new routines, including an XML algorithm, to convert the trained model into a format that humans can easily read, such as an Orange Data Mining export. By incorporating additional statistical methods, this research study strived to achieve higher accuracy in its statistical analysis, surpassing the traditional methods of using Cronbach's alpha and the SEM. Consequently, the study does not adhere to the statistical methods of classical test theory (CTT) due to the increased precision offered by the additional techniques. The statistical methods employed in this study surpassed those of CTT because the IELTS test results of the candidates suggested a potential underlying relationship between the four IELTS macro skills. Upon visually examining the data, it was found that the skills may not be equally weighted as claimed by IELTS (British Council, 2021; IELTS, 2019b, 2021b), prompting the use of more advanced statistical techniques.

Nevertheless, the statistical models could be reproduced and operated with an accuracy of 99.9%. Nagle et al. (2022) understood that artificial intelligence approaches such as ML, data mining, and deep learning, supported by artificial intelligence systems (see

Fig.) such as decision trees, artificial neural networks, support vector machines, and artificial neural networks, could also demonstrate a relationship between dependent variables to a high level of accuracy. Therefore, it is anticipated that using the random forest would deliver precise results to more than 95%.

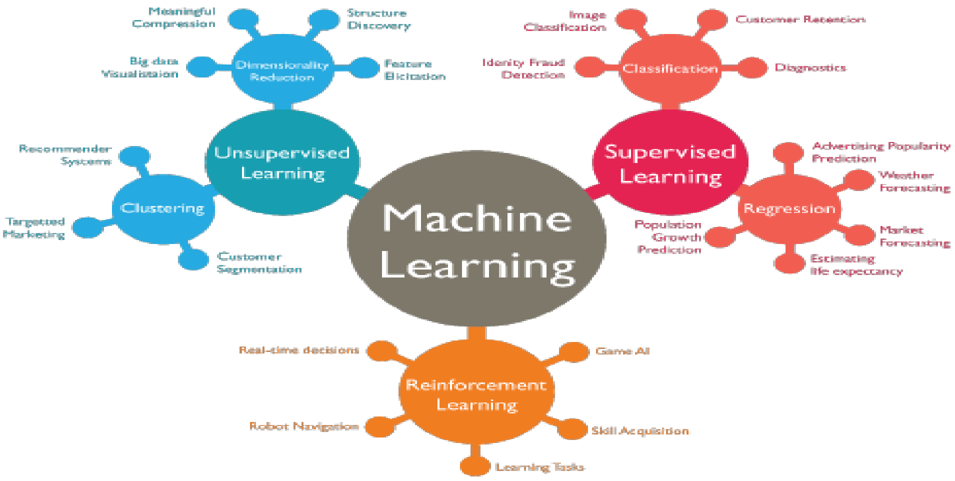


Fig. 2: Types of machine learning algorithms (Nagle et al., 2022, 8)

Furthermore, this study utilised the Orange Data Mining program to analyse biased and unbiased data categories of the Writing macro skill to generate desired outputs, prediction models, and forecasts to address the researchers' concerns. A slider was implemented to facilitate a sensitivity analysis of the number of IELTS test candidates who could achieve a passing score. The percentage of successful test candidates was translated into a probability index, the odds ratio, which ranged between 1 and 0. By adjusting the slider for each of the four IELTS macro skills, the odds ratio could be altered from a lower to a higher IELTS test band score. An odds ratio of 1 indicates the probability of obtaining a pass-or-fail outcome. Coefficients less than 1 indicate a lower level of predictability for the target variable, which in this study was designated as a pass-or-fail outcome.

3.3 Neural Network and Logistic Regression

Both a neural network and a logistic regression model were developed to handle the IELTS band scores, which were considered categorical data. The logistic regression model was used to assess the contributions of the different IELTS test band scores towards the pass/fail outcome, as it was believed that, according to IELTS, they should be equally weighted (British Council, 2021; IELTS, 2019b, 2021b). The Orange Data Mining program's visual programming tool was used to create the workflow depicted in Figure 3.3, which addresses the inquiry of how much the IELTS test band scores add to the prediction value of the overall score.

The statistical technique of "2-fold cross-validation" was used to develop highly accurate models, which involves randomly shuffling the dataset into two sets of equal size, labelled d0 and d1. The two sets are usually created by mixing the data array and splitting it in two. The 2-fold cross-validation used in this study achieved a high accuracy of 0.999/1.0 for both sets, allowing for predicting the "passed/failed" target variable based on the IELTS test band scores for the four macro skills.

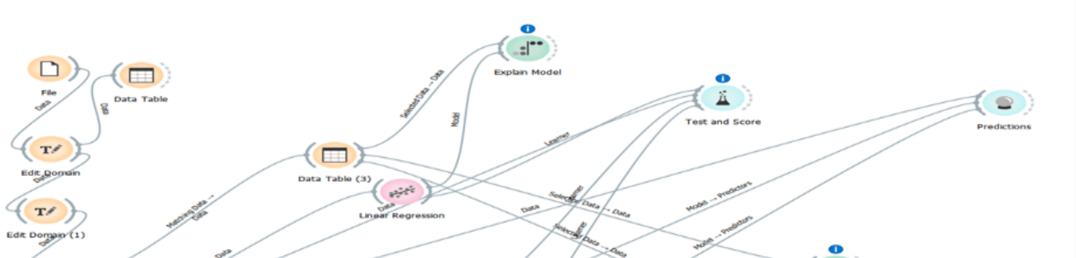


Fig. 3: The Orange Data Workflow

Fig. 3 displays the outcome of the overall IELTS test macro skills prediction model using the Orange Data Mining program. Even though the providers of the IELTS test claim that all four macro skills are equally weighted, this study tested four models, where the overall IELTS test results were the target variable and the input variable was LRWS (listening, reading, writing, and speaking). The study calculated the expected predictions assuming that the four macro skills were equally weighted.

4 Findings and Analysis

IELTS test providers promote their test as a reliable, valid, and secure test of real-life English communication abilities for education, immigration, and professional accreditation (IELTS, 2019c). They have claimed that the IELTS test is reliable, but the definition of reliability is nuanced.

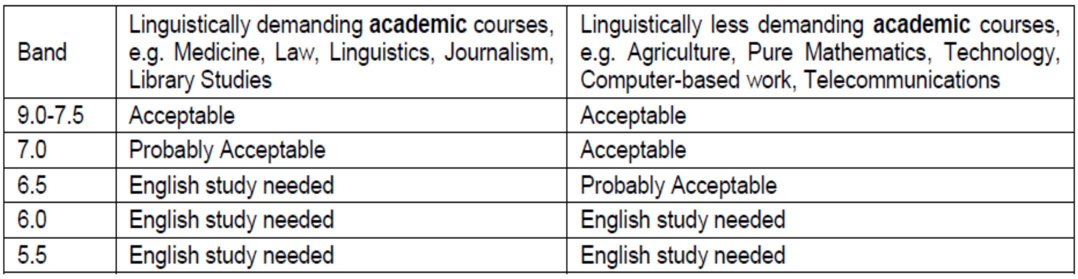
4.1 Standard Error of Measurement (SEM) and Confidence Intervals for Test Retakers

The purpose of this study is not to determine the correct standard error of measurement for different test samples but rather to highlight the fact that there is no consensus on how to measure the reliability of the IELTS test precisely. Elder et al. (2016) suggested that a measurement error of over 0.7 for low-stakes tests is acceptable, while for high-stakes tests, a level higher than 0.8 or 0.9 is acceptable. Wilde (2002) suggested a reasonable level of 0.85 or higher, but Weir (2005) argued that a reliability estimate of greater than 0.9 is desirable for an important test. Since the IELTS test is considered a high-stakes test, it should have a measurement error level of 0.9, not between 0.8 and 0.9, as claimed by the IELTS test providers (IDP: IELTS Australia, 2011). However, for this study of a high-stakes test, confidence levels should be set at 95%, but this requires the IELTS test users to demand it.

Lower confidence levels may be beneficial for the IELTS test providers, as this would allow them to claim a higher level of reliability than is the case. However, users of the IELTS test may not be aware of the statistical implications of the test and would likely rely on the recommendations provided by the IELTS test providers regarding acceptable test scores (refer to Tables (1, 2 and 3).

O'Loughlin (2015) and Spolsky (1997), cited in O'Loughlin (2015), suggested that, as a matter of fairness, when cut-off scores are calculated, they should allow for the SEM of the test. For example, O'Loughlin (2015) pointed out that when the SEM of the overall IELTS test band scores are less than 0.5, and an IELTS test candidate obtains a score of Band 6.5, the actual score obtained falls within a range of Band 6.0 to Band 7.0.

Table (1): IELTS guidance on acceptable language proficiency levels for different academic courses (Ingram & Bayliss, 2007)



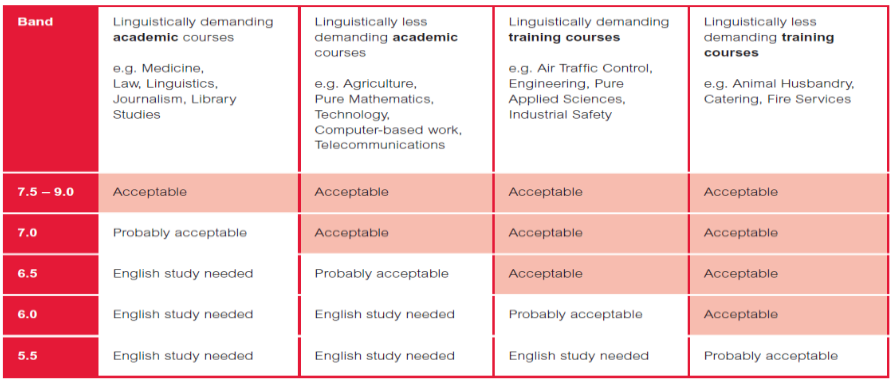
Band	Linguistically demanding **academic** courses, e.g. Medicine, Law, Linguistics, Journalism, Library Studies	Linguistically less demanding **academic** courses, e.g. Agriculture, Pure Mathematics, Technology, Computer-based work, Telecommunications
9.0-7.5	Acceptable	Acceptable
7.0	Probably Acceptable	Acceptable
6.5	English study needed	Probably Acceptable
6.0	English study needed	English study needed
5.5	English study needed	English study needed

Table (2): Acceptable IELTS band requirements for various programs (IELTS, 2013,13; 2016b,13)

Band Score	Linguistically Demanding Academic Courses	Linguistically Less Demanding Academic Courses	Linguistically Demanding Training Courses	Linguistically Less Demanding Training Courses
7.5 – 9	Acceptable	Acceptable	Acceptable	Acceptable
7.0	Probably Acceptable	Acceptable	Acceptable	Acceptable
6.5	English Study Needed	Probably Acceptable	Acceptable	Acceptable

6.0	English Study Needed	English Study Needed	Probably Acceptable	Acceptable
5.5	English Study Needed	English Study Needed	English Study Needed	Probably Acceptable

Table (3): Acceptable IELTS test band scores for Academic and General Training Courses



Band	Linguistically demanding **academic** courses e.g. Medicine, Law, Linguistics, Journalism, Library Studies	Linguistically less demanding **academic** courses e.g. Agriculture, Pure Mathematics, Technology, Computer-based work, Telecommunications	Linguistically demanding **training courses** e.g. Air Traffic Control, Engineering, Pure Applied Sciences, Industrial Safety	Linguistically less demanding **training courses** e.g. Animal Husbandry, Catering, Fire Services
7.5 – 9.0	Acceptable	Acceptable	Acceptable	Acceptable
7.0	Probably acceptable	Acceptable	Acceptable	Acceptable
6.5	English study needed	Probably acceptable	Acceptable	Acceptable
6.0	English study needed	English study needed	Probably acceptable	Acceptable
5.5	English study needed	English study needed	English study needed	Probably acceptable

The IELTS test providers have published the 2015 test performance of IELTS test takers on their website (IELTS, 2016a). The website lists the test version and Cronbach's alpha values for the reading and listening modules for both the Academic and General Training versions. However, some values have been redacted, and the overall average of Cronbach's alpha is reported to be 0.91. Yet, the IELTS test providers ceased publishing any further analytic information has been published about Cronbach's alpha and standard error of measurement values on the IELTS website for their annual IELTS test-takers' performance for 2018 and beyond.

Table (4): Reliability of Reading and Listening modules (IELTS, 2016 a)

Module (All Academic and General Training versions)		**alpha**
Listening version	741	0.911
Listening version	742	0.893
Listening version	743	0.881
(Part of the table was excised here)		
Average alpha across versions		**0.91**

In the earlier years, when the IELTS test providers reported the test performances of the IELTS test, the General Training and Academic reading macro skill versions were regularly reported. The General Training reading macro skill hasan average alpha of 0.92 across all versions with an upper limit of 0.934 and a lower limit of 0.899. The Academic reading, with an average alpha of 0.90 across all versions, has an upper limit of 0.92 and a lower limit of 0.80. For a high-stakes assessment test, it would be part of the test validity to have all test stakeholders agree on what alpha level should be the cut-off point. Should it be 0.80 or 0.90 or 0.95? For the validity of the assessment test, there must be an agreement by all test stakeholders on what would be the appropriate alpha to use.

Table (5): Mean, standard deviation, and standard error of measurement of Listening and Reading (IELTS, 2016 a)

Module	Mean	SD	alpha	SEM
Listening	6.10	1.3	0.92	0.37
ACR	6.02	1.2	0.90	0.38
GTR	6.00	1.5	0.91	0.45

The IELTS test providers have excluded the writing and speaking macro skills from Table (6) as they are not based on individual test items but on criteria. Although the table includes information on SEM, the IELTS test providers do not reference it beyond that. Based on the available data, it can be inferred that the SEM measurements for the listening and reading macro skills in the IELTS test are too high to support the test's reliability, especially given its high-stakes nature.

Table (6): How your overall IELTS test band scores are calculated (IELTS, 2018)

	Listening	Reading	Writing	Speaking	Average of four components (total of the four individual component scores divided by four	Band score
Test-taker A	6.5	6.5	5	7	6.25	6.5
Test-taker B	4.0	3.5	4.0	4.0	3.875	4.0
Test-taker C	6.5	6.5	5.5	6.0	6.125	6.0

Moreover, there is uncertainty regarding the accuracy of IELTS test band scores awarded to test candidates. The IELTS test partners annually report their annual test performances (IELTS, 2019c) that their test has a standard error of measurement (SEM), but they do not provide further details about what this means. It is unclear whether the SEM is an average of the standard error of measurement for all tests administered during the year or whether it applies to each test.

The standard error of measurement is directly related to the reliability of a test. That is, the larger the standard error of measurement, the lower the reliability of the test and the less precision there is in the measures taken and scores obtained. Since all measurement contains some error, it is highly unlikely that any test will yield the same scores for a given person each time they are retested (Bishop, 1996).

The interpretations given are subject to "confidence levels" (Bulpitt, 1987). The confidence level has an upper and lower limit, covering the true but unknown score. The confidence that one has in the true score being included is based on the empirical rule, also known as the three-sigma rule or the 68-95-99.7 rule (Maronna et al., 2019). Another way to explain it is if the trials had a 95% confidence interval and the test was repeated 1,000 times, this confidence level should include the true but unknown value 950 times.

However, because of the influence of the IELTS test on determining the future of test candidates––and the IELTS test is seen as a high-stakes test––the confidence level should be set at 95%, and when the IELTS test band scores are reported, so should the standard error of measurement be included. Then the following would provide much more information to the IELTS test users. In other words, while the test results may demonstrate high reliability, it is important to note that the analysis is based on IELTS test band scores that were altered at various points and are not raw scores being evaluated for reliability.

If the SEM = 0.37, then the range for 95% confidence is +/- 1.96 X 0.37 = +/- 0.7252.

Therefore, if a candidate is given a raw score of 6.10, one can be very confident that his or her score lies between 6.10 +/- 0.252, that is, 5.3748 to 6.8252––a grade of between Band 5.5 and 7.0.

The suggestion is that providing IELTS test band scores to candidates without clarifying the relevant test standard error of measurement can be misleading for both the candidates and test users. The analysis in Tables (7) and (8) involved examining the test candidates' subsequent sittings after their initial one, which was not included in the analysis. This approach was taken to identify any discernible patterns that emerged.

Table (7) provides the reader with insight into the number of test candidates who retake the IELTS test more than once and to determine whether they are impacted by the "test familiarity" or "practice effect", which could potentially influence their final IELTS test band scores. This information can also help better understand the statement made by Warwick Freeland, managing director of IELTS at IDP Education, who claimed that most test-takers only sit IELTS once (cited in Mina, 2019).

Table (7): Analysis of the test-retakers by test centre

Location	Number of test Candidates	Number of test Retakers	Percentage of test Retakers	Took the test 2 x	Took the test 3 x	Took the test 4 x	Took the test 5 x	Took the test 6 x	Took the test 7 x	Absent	Refund
Canberra	1,023	99	9.7%	87	11	1				23	4
Hobart	886	182	20.5%	126	40	14	0	1	1	69	18
Launceston	458	65	14.2%	31	16	14	1	2	1	4	1
Iran	4763	132	2.8%	129	3	-	-	-	-	14	2
Total	7,130	478	18.88%	373	70	29	1	3	2	110	25

Table (8) summarises the reliability measures, specifically Cronbach's alpha (α), computed for the four IELTS macro skills: listening, reading, writing, and speaking, categorized by the number of tests taken. The values presented in the table represent the measure of reliability for each macro skill. The first set of values in the table indicates the analysis condition based on the number of items or variables included in the analysis. Similar to Table (5), the IELTS test providers do not include the writing and speaking macro skills in this analysis because they are criterion-referenced rather than item-based. As a result, Table 4.8 only includes reliability measures for the listening and reading macro skills.

The scores of the candidates who retook the IELTS test, as presented in Table (8), do not include their first attempt. Although the reliability measure of test-retest, which measures test consistency over time, could have been used to include the first attempt, this would have served the same purpose as the current analys: to observe the value of Cronbach's alpha. However, this analysis does not measure the consistency of responses to the items between each of the four IELTS macro skills in the test and may, therefore, not be an effective test construct to measure the candidates' English language proficiency skills. Instead, it may only indicate their English language competency.

Table (8): Academic IELTS test retakers' scores

Number of Times Test Taken	Item	Number of Valid Cases (N)	Reliability Statistic (Cronbach's alpha)
Two	Listening	477	.884
	Reading	477	.854
Three	Listening	101	.888
	Reading	101	.871
Four	Listening	31	.907
	Reading	31	.753
Five	Listening	7	Sample too small

	Reading	7	Sample too small
Six	Listening	4	Sample too small
	Reading	4	Sample too small
Seven	Listening	2	Sample too small
	Reading	2	Sample too small
Eight	Listening	1	Not Computed
	Reading	1	Not Computed

The final set of values presented in Table (8) represents the measure of reliability, as indicated by Cronbach's alpha (α), for the four IELTS macro skills. The initial set of values in the table represents the analysis condition based on the number of items or variables included. Each item represents a list of scores obtained in a separate sequence. This table can be presented graphically to understand better the test reliability for each of the four IELTS test bands. It shows that for a high-stakes test, one would expect to see an alpha of 0.95 or higher. The range of values presented in Table 4.8 for the four macro skills is between 0.753 and 0.979, with the highest alpha value indicating almost perfect test reliability.

Table (9): Statistical analysis for the reliability of the IELTS test results of test candidates

		Listening	Reading	Writing	Speaking	Overall Score
N	Valid	767	767	767	767	767
	Missing	0	0	0	0	0
Mean		6.0	5.8	5.5	5.9	5.9
Median		6.0	5.5	5.5	6.0	6.0
Std. Deviation		1.1321	1.0288	.8620	1.0151	.8474
Skewness		.016	.457	-.342	-.670	.008
Std. Error of Skewness		.088	.088	.088	.088	.088
Kurtosis		.354	.272	1.285	3.869	.347
Std. Error of Kurtosis		.176	.176	.176	.176	.176
Percentiles	25	5.0	5.0	5.0	5.5	5.5
	50	6.0	5.5	5.5	6.0	6.0
	75	6.5	6.5	6.0	6.5	6.5

Table (9) and Table (10) were examined to gain insights into the IELTS test band scores found in the study's database. It revealed that the mean scores, when rounded to the nearest IELTS test band score, tended to cluster around the same values as those found in Table (11) and Table (12). The latter represents the various measures of central tendency, dispersion, and shape of the IELTS test scores. Since the variables (scores) are ordinal, the median may be considered the primary measure of central tendency.

Table (10): Separate statistical measurements of 1,121 Test candidates

		Listening	Reading	Writing	Speaking	Overall Score
N	Valid	1121	1121	1121	1121	1121
	Missing	0	0	0	0	0
Mean		5.8	5.6	5.6	6.1	5.8
Median		6.0	5.5	5.5	6.0	6.0
Std. Deviation		1.0854	1.0588	.8395	.9522	.7963
Skewness		.134	.334	-.414	-.798	-.012
Std. Error of Skewness		.073	.073	.073	.073	.073
Kurtosis		.339	.432	1.217	3.949	.473
Std. Error of Kurtosis		.146	.146	.146	.146	.146
Percentiles	25	5.0	5.0	5.0	5.5	5.5
	50	6.0	5.5	5.5	6.0	6.0
	75	6.5	6.0	6.0	6.5	6.5

It is important to note that Table (11) included all the candidates who took the IELTS test only once. One significant observation from the results is that the IELTS test appears reliable.

Table (11): Distribution for the Standard deviation and the Standard error of the mean of first-time test takers for the Academic and General Training IELTS tests

Listening Reading Writing Speaking					
		Listening	Reading	Writing	Speaking
Academic	Mean	6.2138	5.8325	5.7383	6.3005
	N	15664	15664	15664	15664
	Std. Deviation	1.17523	1.23064	.92397	.97336
	Std. Error of Mean	.00939	.00983	.00738	.00778
General Training	Mean	6.0671	5.6827	5.8417	6.3417
	N	10878	10878	10878	10878
	Std. Deviation	1.21671	1.25734	.88267	.94340

	Std. Error of Mean	.01167	.01206	.00846	.00905
Total	Mean	6.1537	5.7711	5.7807	6.3174
	N	26542	26542	26542	26542
	Std. Deviation	1.19456	1.24382	.90868	.96139
	Std. Error of Mean	.00733	.00763	.00558	.00590

The reliability of the IELTS test was found to be consistent across different groups of test takers when comparing two different IELTS tests. This noticed was in both Table (12), which included all test takers who took the IELTS test only once, and Table (13), which included those who took the test two or more times. In both cases, the IELTS test was found to be reliable.

Table (12): Overall reliability of the two IELTS tests

	Mean	N	Std. Deviation	Std. Error of Mean
Academic Test	6.0860	15664	.93031	.00743
General Training Test	6.0448	10878	.91317	.00876
Total	**6.0691**	**26542**	**.92353**	**.00567**

Upon further examination of the dataset used in the current study, reliability statistics were calculated and reported, and the standard error of measurement was used to construct 68% and 95% confidence intervals for the four test band scores. This information is presented in Table (13). The inclusion of the standard error of measurement and confidence intervals differs from the information presented in Table (13) and Table (14), which focused primarily on the reliability of the IELTS test.

Furthermore, Table (13) illustrates that IELTS test band scores are not exact but fall within a range due to the standard error of measurement and confidence levels. The table shows the standard error of measurement for each of the four IELTS macro skills, as well as the 68% and 95% confidence intervals for each test band score. The upper and lower limits of the confidence intervals indicate the range in which a test candidate's true score will likely fall, depending on the level of certainty or confidence used in the analysis. This observation suggests that IELTS test band scores are not precise but rather lie within a range, with the 95% confidence interval providing a more precise figure.

Table (13): The 68% and 95% confidence intervals of band test scores based on SEM and reliability statistics

Band	Retake	No. of Candidates	Cronbach's alpha	Mean	SD	SEM	68% Confidence Interval		95% Confidence Interval	
							Lower Limit	Upper Limit	Lower Limit	Upper Limit
Listening	2	2677	0.632	6.35	1.179	0.715	5.64	7.07	4.92	7.78
	3	330	0.091	6.18	1.216	1.159	5.02	7.34	3.86	8.50
	4	90	0.036	6.24	1.171	1.150	5.09	7.39	3.94	8.54
Reading	2	2677	0.667	5.92	1.211	0.699	5.22	6.62	4.52	7.32
	3	330	0.141	5.89	1.186	1.100	4.79	6.99	3.69	8.09
	4	90	0.265	5.89	1.116	0.957	4.93	6.84	3.97	7.80
Writing	2	2677	0.671	5.78	0.947	0.544	5.23	6.32	4.69	6.87
	3	330	0.032	5.70	0.983	0.967	4.73	6.67	3.77	7.63
	4	90	0.184	5.72	1.034	0.934	4.78	6.65	3.85	7.58
Speaking	2	2677	0.664	6.26	1.029	0.596	5.66	6.86	5.07	7.45
	3	330	0.044	6.16	1.062	1.039	5.12	7.20	4.09	8.24
	4	90	0.169	6.18	1.077	0.982	5.20	7.16	4.22	8.15

Where N = Total Number of Entries, n = Total Number of Entries per Country, SEM = Standard Error of Mean, Mdn = Median, SD = Standard Deviation, Skew = Skewness Measure, SES = Standard Error of Skewness, Kurt = Kurtosis Measure, SEK = Standard Error of Kurtosis

Table (14) presents a descriptive summary of the dataset used in this study, which consists of valid and cleaned data from 2,677 candidates. The table provides statistical analysis for the four macro skills of the IELTS test, including measures such as minimum and maximum band scores, mean and standard error of the mean, median, standard deviation, skewness, and kurtosis. While this analysis indicates a high reliability of the test results, it is important to note that the IELTS test band scores used were adjusted at various stages and are not raw scores being tested for reliability.

Table (14): Descriptive statistics of IELTS test scores across countries

Country	Band	Mean	the standard error of Mean	68% confidence interval for the Mean		Mdn	SD	95% confidence interval for Mean		Min	Max	Skew	Kurt
				Lower Limit	Upper Limit			Lower Limit	Upper Limit				
Australia	Listening	6.43	.016	6.395	6.457	6.50	1.145	4.137	8.715	1.00	9.00	-.055	.121
	Reading	5.95	.017	5.916	5.981	6.00	1.206	3.536	8.361	1.00	9.00	.170	.337
	Writing	5.79	.013	5.766	5.817	6.00	.934	3.924	7.659	.50	9.00	-.155	2.321
	Speaking	6.28	.014	6.250	6.306	6.00	1.031	4.216	8.339	1.00	9.00	.001	1.033
China	Listening	6.41	.139	6.132	6.693	6.50	.876	4.660	8.165	4.50	8.00	-.020	-.733
	Reading	6.46	.127	6.205	6.720	6.50	.804	4.855	8.070	5.00	8.50	.752	1.002
	Writing	5.95	.103	5.743	6.157	6.00	.648	4.653	7.247	5.00	7.50	.269	-.595
	Speaking	6.19	.155	5.875	6.500	6.25	.979	4.230	8.145	4.00	7.50	-.472	-.487
India	Listening	6.57	.129	6.313	6.830	6.50	.901	4.769	8.374	4.50	8.50	.400	-.076
	Reading	5.65	.119	5.415	5.892	5.50	.830	3.992	7.314	4.00	7.50	.000	-.249
	Writing	5.73	.101	5.531	5.938	5.50	.708	4.319	7.150	4.00	7.00	.157	-.100
	Speaking	6.46	.131	6.196	6.723	6.50	.918	4.624	8.294	5.00	8.50	.357	-.582
Iran	Listening	5.38	.082	5.216	5.541	5.50	1.122	3.135	7.622	1.00	8.50	-.430	1.951
	Reading	5.52	.089	5.344	5.694	5.50	1.205	3.109	7.929	2.00	9.00	-.043	-.116
	Writing	5.78	.076	5.635	5.933	6.00	1.028	3.728	7.839	2.00	8.00	-.865	1.489
	Speaking	6.13	.068	5.992	6.262	6.00	.927	4.272	7.982	3.00	9.00	-.357	1.552
Kazakhstan	Listening	5.01	.096	4.820	5.198	5.00	1.257	2.494	7.523	1.00	8.00	-.447	.435
	Reading	5.53	.081	5.370	5.688	5.50	1.056	3.416	7.642	1.00	8.00	-.224	1.017
	Writing	5.30	.095	5.108	5.485	5.50	1.250	2.796	7.797	.50	7.50	-1.464	2.830
	Speaking	5.69	.088	5.515	5.863	5.50	1.159	3.371	8.006	2.00	8.00	-.742	.709
USA	Listening	5.17	.219	4.728	5.613	5.00	1.455	2.261	8.079	2.00	8.50	.577	.714
	Reading	5.17	.206	4.756	5.585	5.00	1.364	2.443	7.898	3.00	9.00	1.064	1.358
	Writing	5.44	.280	4.879	6.007	5.50	1.856	1.731	9.155	2.00	9.00	.151	-.220
	Speaking	6.20	.204	5.793	6.616	6.00	1.353	3.500	8.910	4.00	9.00	.496	-.306
Other	Listening	6.32	.171	5.978	6.664	6.50	1.245	3.832	8.810	3.50	9.00	-.179	-.603
	Reading	6.20	.180	5.837	6.559	6.00	1.310	3.579	8.817	4.00	9.00	.356	-.797
	Writing	5.58	.085	5.415	5.755	5.50	.618	4.349	6.821	4.50	7.00	.235	-.371
	Speaking	6.13	.118	5.896	6.368	6.00	.856	4.421	7.843	4.00	8.00	-.142	.104
Total	Listening	6.340	.016	6.311	6.372	6.50	1.189	3.962	8.718	1.00	9.00	-.159	.345
	Reading	5.920	.016	5.889	5.951	6.00	1.206	3.508	8.332	1.00	9.00	.163	.339
	Writing	5.770	.013	5.748	5.797	6.00	.956	3.858	7.682	.50	9.00	-.307	2.618
	Speaking	6.250	.014	6.228	6.281	6.00	1.037	4.176	8.324	1.00	9.00	-.046	1.070

Where N = Total Number of Entries, n = Total Number of Entries per Country, SEM = Standard Error of Mean, Mdn = Median, SD = Standard Deviation, Skew = Skewness Measure, SES = Standard Error of Skewness, Kurt = Kurtosis Measure, SEK = Standard Error of Kurtosis

Based on Table 15, it can be observed that the standard error of measurement (SEM) varies according to the location of the test centre, with the USA having the highest SEM followed by China. However, it should be noted that the sample sizes were small, with statistical analysis only being conducted for countries with a sample size of over 50 participants. The SEM is directly related to the test's reliability, and a smaller SEM indicates a more reliable reflection of the total population.

Table (15): Distribution by the country for the Standard deviation and the Standard error of the mean

Countries with more than 50 members in the test sample		Listening	Reading	Writing	Speaking
Australia	Mean	6.3864	5.8841	5.7685	6.2831
	N	17849	17849	17849	17849
	Std. Deviation	1.14303	1.23018	.88165	.98228
	Std. Error of Mean	.00856	.00921	.00660	.00735
China	Mean	6.1827	6.0962	5.7981	5.8942
	N	52	52	52	52
	Std. Deviation	1.03388	1.10719	.70196	1.09963
	Std. Error of Mean	.14337	.15354	.09734	.15249
India	Mean	6.6639	6.0984	6.0328	6.4590
	N	122	122	122	122
	Std. Deviation	1.08224	1.17417	.89705	1.02165

	Std. Error of Mean	.09798	.10630	.08122	.09250
Iran	Mean	5.5687	5.3209	5.9395	6.4824
	N	4646	4646	4646	4646
	Std. Deviation	1.08855	1.18003	.80839	.80600
	Std. Error of Mean	.01597	.01731	.01186	.01182

Khazakstan	Mean	5.0355	5.5628	5.3361	5.7350
	N	183	183	183	183
	Std. Deviation	1.25689	1.08695	1.24053	1.17645
	Std. Error of Mean	.09291	.08035	.09170	.08697
U.K.	Mean	5.7725	5.7648	5.6331	6.2942
	N	3505	3505	3505	3505
	Std. Deviation	1.16746	1.25243	1.07259	.97869
	Std. Error of Mean	.01972	.02115	.01812	.01653
USA	Mean	5.1889	5.1889	5.4333	6.2000
	N	45	45	45	45
	Std. Deviation	1.44320	1.35382	1.83588	1.33740
	Std. Error of Mean	.21514	.20182	.27368	.19937
Total	Mean	6.1504	5.7671	5.7783	6.3157
	N	26402	26402	26402	26402
	Std. Deviation	1.19247	1.24177	.90737	.96012
	Std. Error of Mean	.00734	.00764	.00558	.00591

More immediate interest in Table 16 is that the mean value appears similar to the other statistical analyses already carried out. One explanation is that the mean scores are being "managed" rather than due to the random results of different test candidates who sit for the IELTS test each week at the different test centres. Table 15 and Table 16 provide some evidence of the good reliability of the IELTS test, as indicated by the relatively small standard errors of measurement. However, it is worth noting that the mean scores tend to be similar across different countries, and it is unclear why this is the case. The IELTS test providers have not explained why the test results tend to be similar when quantized by the IELTS quantizing rules.

Table (16): Test reliability by country

	Mean	N	Std. Deviation	Std. Error of Mean
Australia	6.1450	17849	.91549	.00685
China	6.0673	52	.82859	.11490
India	6.4180	122	.90991	.08238
Iran	5.8876	4646	.83258	.01221
Khazakstan	5.4754	183	1.06037	.07839
UK	5.9260	3505	.98484	.01663
USA	5.5889	45	1.38289	.20615
Total	**6.0662**	**26402**	**.92149**	**.00567**

In order to obtain further evidence for answering the research question and verifying the claim of IELTS test providers regarding its reliability, a cluster analysis was conducted.

4.2 The Cluster Analysis (The Doughnut)

The Orange Data Workflow (depicted in Fig. 4) was utilised, and a Neural Network model was trained with data imported from the study's database. This model targeted the distributions of Band 4 and Band 7.

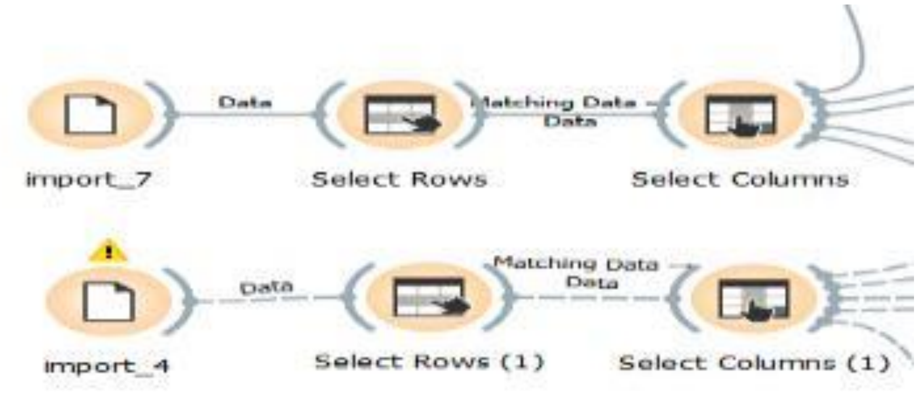


Fig. 4: The Orange Data Workflow

The resulting workflow can be observed in Figure 5, which displays the Orange Data Workflow for the Neural Network model.

After preparing the Orange Data Workflow, the next step in the analysis was to train a Neural Network model using the data imported from the study's database. The purpose of this model was to predict the distributions of Band 4 and Band 7, which were set as the target variables.

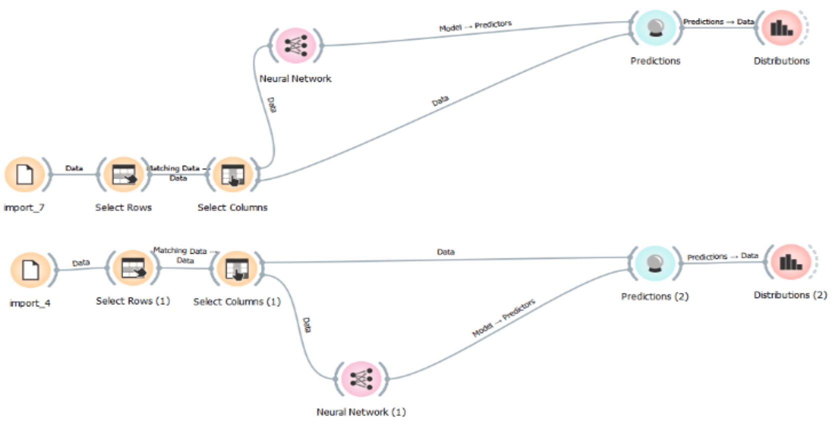


Fig. 5: The Orange Data Workflow for the Neural Network Model

After developing the Neural Network model, the exploratory analysis proceeded to add a prediction module, which allowed for predictions based on the model. For example, the following graph (refer to Fig. 6) illustrates the results of the Neural Network model, with the x-axis representing Band 7 and the y-axis representing the number of test candidates who passed or failed. Fig.6 and Fig. 7 show remarkable accuracy in their predictions. In particular, for Band 7, the prediction model achieves a perfect 100% accuracy. For Band 4, the accuracy is 99%, meaning that only around 1% of predictions are inaccurate.

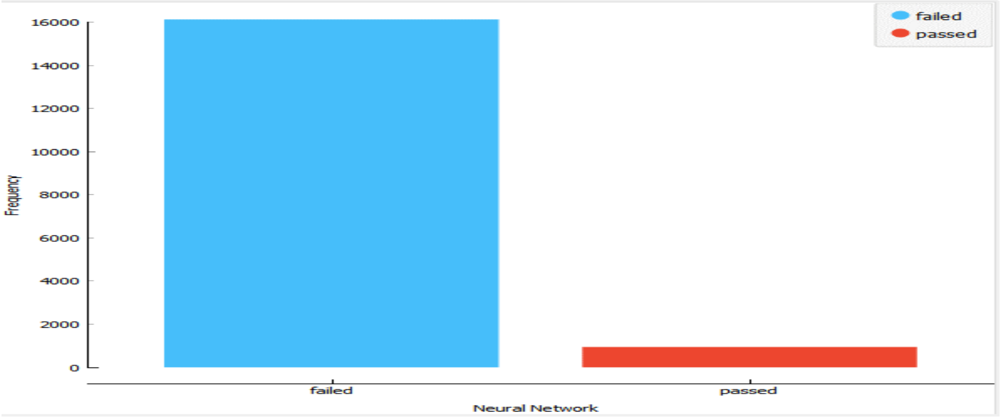


Fig. 6: Analysing the performance of prediction using machine learning classification algorithms for Band 7

Fig. 7 displays information regarding the total number of IELTS test candidates who either passed or failed Band 4, with the x-axis representing the different values of passed and failed candidates.

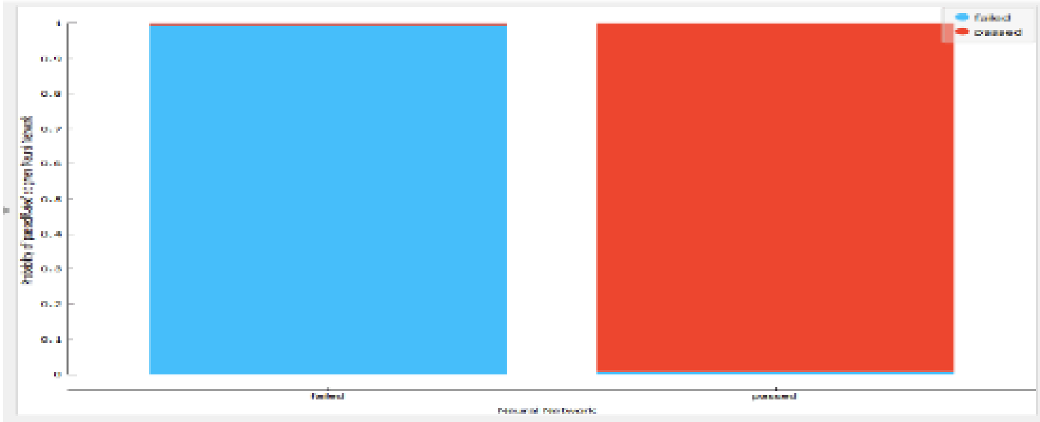


Fig. Error! Reference source not found.**7:** the total number of test candidates who passed/failed Band 4 on the x-axis

Fig. 8 illustrates the Orange Data Workflow specifically designed for the training of Band 7. The workflow consists of several components, including data input, data pre-processing, feature engineering, model selection, hyperparameter tuning, and model evaluation. The input data is first pre-processed to remove any missing values or outliers, and feature engineering is performed to extract relevant features from the data. Next, different models are selected based on performance, and hyperparameters are tuned to optimize accuracy. Finally, the best model is evaluated on the test data to ensure its effectiveness. This workflow provides a structured and efficient approach to training a model for Band 7 using the IELTS test data.

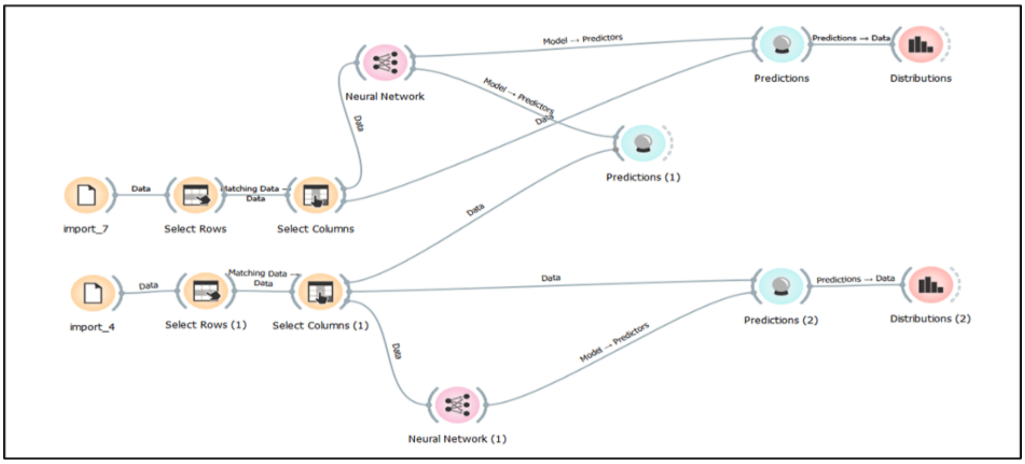


Fig. 8: The Orange Data Workflow for Training Band 7

4.3 The Linear Regression Results

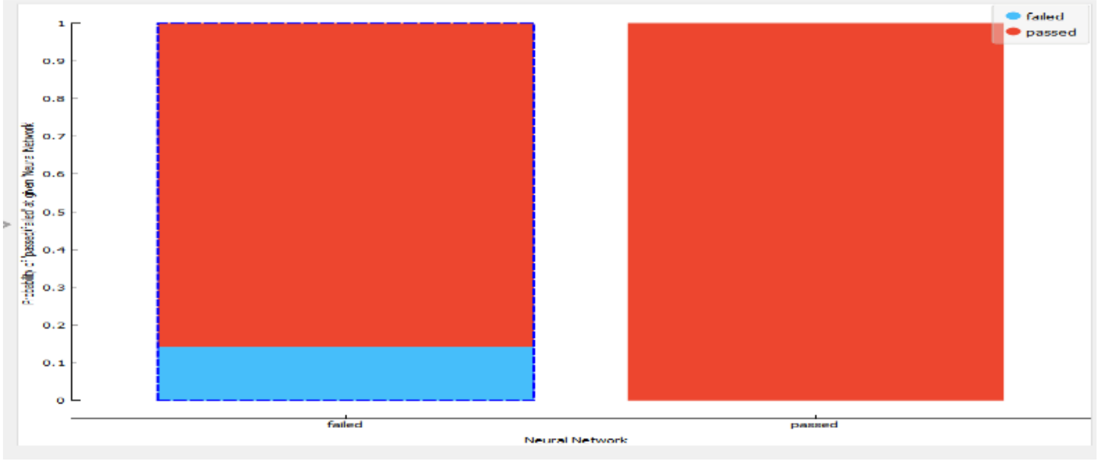


Fig. 9: Analysing the Performance of Prediction using Machine Learning (ML) Classification Algorithms for Band 4

The following example is a reverse of the previous one. The model used is to compare the probability of passed/failed for Band 4 with the data of Band 7:

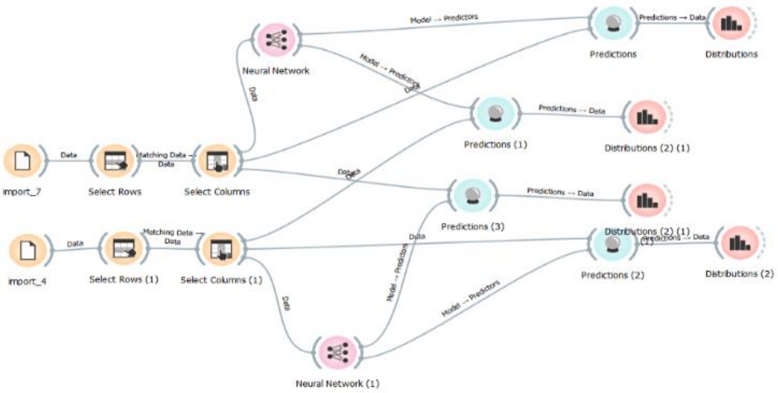


Fig. 10: The Orange Data Workflow for Comparison for Passed/Failed for Band 4 with Data for Band 7

This study acknowledges the inaccuracy introduced in the prediction model, as finding a completely accurate model is impossible. Instead, the study aims to find a model that provides a plausible approximation of reality and offers insight. The model's inaccuracy is acceptable since IELTS test band scores are generally approximate due to the IELTS rounding procedures. As per Browne (2000), this study employs a cross-validation model to assess the predictive validity of linear regression equations used to forecast the "performance criterion" on test scores obtained from several tests. The study's prediction model is trained with the Band 7 data to predict "passed/failed" with the Band 4 data and vice versa with a range of predictions. For example, the study will start with a model of 7 predicted with data of 4, as shown in Fig. 11.

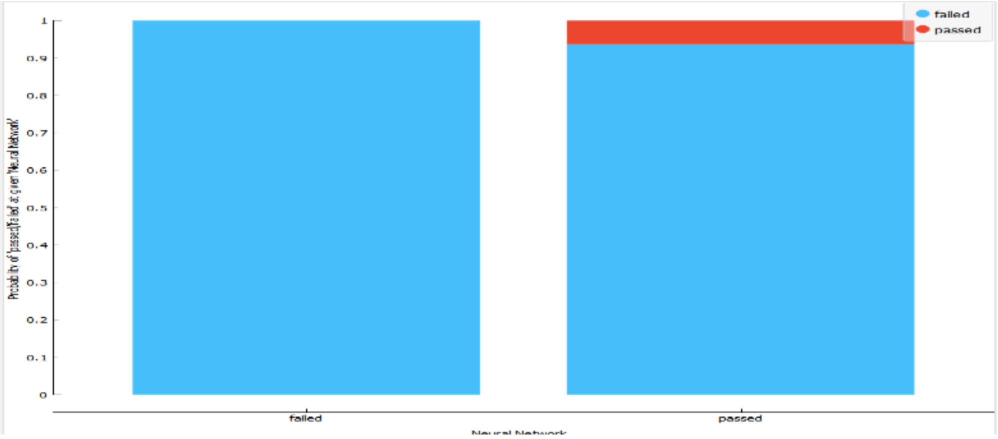


Fig. 11: Analysing the performance of prediction using ML classification algorithms for Band 7

Based on the tables presented so far, it is impossible to accurately determine the predictive validity for candidates who pass the IELTS test. Moreover, the model used for Band 4 cannot be used to predict Band 7 results and vice versa. The analysis used the passed/failed rule set for Band 4, which sets the four IELTS macro skills at Band 4. However, the results would be inaccurate if the same model were used with Band 7 as a pass mark. This finding was unexpected since the rule set treats all IELTS test band scores as being equally weighted. The conclusion is that there may have been some form of interference in the results, which raises questions about the fairness of the test as well.

The reason for this observation may be found in Table 4.23, which shows how the rules vary across different sub-categories, such as the Centre and Retaker categories, in the horizontal view. In contrast, the vertical view highlights differences between sub-categories and the overall model for each module. If the test were truly "fair", there would be minimal differences between sub-categories. However, significant differences are present for each target score and across sub-categories, contradicting the claim that the test is fair regardless of the reason it is taken or where it is taken. The notion of "fair" in this context refers to consistency in scoring and rules across different sub-categories, which is not observed in the results. The observation of Table 4.23 yields the following: When reading, writing, and speaking are on Band 7, the probability of passing the listening test is 95%, reading is 75%, the writing test is 51%, and the speaking test is 82%.

The sensitivity tests could reveal that the probability of passing or failing in the writing macro skill at band 4 is 50%, but when it comes to band 5, the reading macro skill becomes critical, and as the band scores for the reading macro skill increase, it has a negative impact on the writing macro skill. The reason for this behaviour can be explained by looking at the table, which shows that the rule sets change for each category tested, such as the IELTS test centre and retakers, in the horizontal view. The vertical view shows the differences in sub-categories for each of the four IELTS macro skills.

Table (17): Probability of passing IELTS test macro skills

LISTENING	READING	WRITING	SPEAKING	TOTAL CANDIDATES
7880	7880	7880	7880	31520
295	733	1306	366	2700
96.39%	91.49%	85.78%	95.56%	92.11%
944	944	944	944	3776
45	304	891	201	1441
95.45%	75.64%	51.44%	82.45%	72.38%

Despite their lack of transparency, the IELTS test providers appeal to users and candidates to trust them (Templer, 2004). The theme of 'trust' is prevalent throughout the IELTS publications, and the IELTS test providers rely on their reputation and prestige instead of using evidence-based research to prove the reliability of the IELTS test. This approach is a fallacy of fact known as the "argument from authority" (Friedmann, 1989), which is used to dismiss objections by appealing to higher authorities such as the British Council and Cambridge English, who claim to have superior knowledge and experience. However, relying solely on their reputation and the fallacy of fact is insufficient to prove the reliability of the IELTS test's design, and independent research is needed to ensure the test's quality. As McNamara (2000) stated, every test is vulnerable to critical inquiry.

In recent years, many organisations in the United States have been self-reflection on the damaging effects of assessment tests. For example, the American Psychological Association [APA] (2021) recently acknowledged their past involvement in promoting systematic racism against people of color. Though it may not have been intentional, the American Psychological Association [APA] (2021) failed to encourage researchers to report the error margin associated with the test scores. Moreover, Perlin et al. (2021, 464-465) noted that the SEM, in the judgement Atkins v Florida (Hall v. Florida, 134 S. Ct. 1986, 1998 (2014)) that the professionals who design, administer, and interpret.Q.I.Q. tests have agreed, for years now, that .Q.I.Q. test scores should be read not as a single fixed number but as a range. The Court stressed: "An individual's intellectual functioning cannot be reduced to a single numerical score." It was an error to use such a test score "without necessary adjustment." As the "vast majority" of states had rejected a strict seventy-point cut-off, and as the trend to recognise the significance of the standard error of measurement was "consisten[t], "this was, to the Court, "strong evidence of consensus that our society could not regard this strict cut-off as proper or humane".

5 Conclusions and Limitations

The IELTS test was originally designed as a tool for universities to diagnose international students' English language proficiency skills and offer them remedial assistance if needed. This dual-purpose approach was highlighted by Davies (2014), who saw the remedial aspect as particularly valuable. However, there were concerns raised by Hirsh (2007) and Winefield et al. (2003) about the costs associated with offering remedial classes, especially given the funding cutbacks universities were facing. Some experts, such as Ingram (2005), viewed using the IELTS test for purposes other than its original design as unethical. Fulcher and Davidson (2007) coined the term "retrofitting" to describe this practice of using an assessment test designed for one purpose for a different purpose.

Furthermore, concerns have been raised about the IELTS test's consistency, with candidates reportedly obtaining varying results each time they take it. The consistency of the test results over multiple administrations determines test reliability. While there are different interpretations of test reliability, this study used a specific definition. However, a systematic review of the literature revealed no evidence of predictive reliability for the IELTS test. Earlier versions of the IELTS test were also found to lack predictive validity properties, as admitted by Alderson and Clapham (1992). The test was designed with face validity in mind, which was deemed sufficient for its validation.

The issue of commercial confidentiality claimed by IELTS test providers poses a challenge to the concept of test design's reliability in this study. According to the American Educational Research Association and American Psychological Association (2014), all stakeholders must be adequately informed of the test's purpose for an assessment test to be considered valid. However, researchers cannot access data and past test papers from IELTS test providers to replicate the test. This limitation raises concerns about the authenticity of the claims made by the company, which cannot be proven.

The limitations of this study increased with the claim that the four macro skills of the IELTS test are equally weighted (British Council, 2021; IELTS, 2019d, 2021b) was not explained by the test providers, which poses a challenge for the interpretation of the results. However, it was consistently observed that the writing macro skill was the most challenging for many IELTS test candidates, as reported in the literature by Hamid (2015), Müller (2015), and Pearson (2019b). These findings suggest that the claim of equal weighting may not accurately reflect the actual difficulty level of each macro skill. The lack of transparency in this regard may affect the test's design fairness and reliability in accurately measuring the test candidates language proficiency.

The IELTS test providers argue that the non-disclosure of information or commercial confidentiality is necessary to safeguard the credibility of the IELTS test (Hogan, 2005; IDP: IELTS Australia, 2011) Yet, there is a broader societal concern about commercial secrecy. It raises questions about how companies should be allowed to restrict access to information that could be used for public benefit.

References

American Psychological Association [APA]. (2021). *Role of Psychology and APA in Dismantling Systemic Racism Against People of Color in U.S.* American Educational Research Association.retrieved from https://www.apa.org/about/policy/dismantling-systemic-racism

Alderson, J. C., & Clapham, C. (1992). Applied linguistics and language testing: A case study of the ELTS test. *Applied Linguistics*, *13*(2), 149-167. https://doi.org/10.1093/applin/13.2.149

American Educational Research Association, & American Psychological Association, a. N. C. o. M. i. E. (2014). *Standards for educational and psychological testing*. https://www.apa.org/science/programs/testing/standards

Arrigoni, E., & Clark, V. (2015). Investigating the appropriateness of IELTS cut-off scores for admissions and placement decisions at an English-medium university in Egypt. *IELTS Research Reports Online Series*, *2015/3*, 29. https://www.ielts.org/en-us/for-researchers/research-reports/online-series-2015-3

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

Bayliss, A., & Ingram, D. E. (2006). *IELTS as a predictor of academic language performance*. In Australian International Education Conference. 1-12. https://pdfs.semanticscholar.org/545d/f30dada9c00f1029b812db812e799ba9e8a8.pdf

Bishop, D. (1996). Standard Error of Measurement. *Technical Assistance Papers*.

Breeze, R., & Miller, P. (2011). Predictive validity of the IELTS listening test as an indicator of student coping ability in Spain. *IELTS Research Reports*, *12*, 201-234. https://search.informit.org/doi/10.3316/informit.16046559270121

British Council, IELTS Australia, & University of Cambridge ESOL Examinations. (2006). *IELTS Brand Guide Book*. Cambridge UP.

British Council. (2021). *IELTS band scores explained.* British Council. https://takeielts.britishcouncil.org/find-out-about-results/understand-your-ielts-scores

British Council. (2023). *IELTS frequently asked questions*. https://www.britishcouncil.org.tr/en/exam/ielts/faq

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108-132. https://doi.org/10.1006/jmps.1999.1279

Bulpitt, C. J. (1987). Confidence intervals. *Lancet*, *1*(8531), 494-497. https://doi.org/10.1016/s0140-6736(87)92100-3

Cambridge ESOL. (2004). IELTS – some frequently asked questions. *Research Notes*, *18*, 14-15.

Cambridge Assessment. (2018). *Research and collaboration.* Cambridge Assessment. https://www.cambridgeenglish.org/research-and-validation/research-and-collaboration/

Cambridge University Press. (2023). *IELTS Test Format*. https://www.cambridgeenglish.org/exams-and-tests/ielts/test-format/

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Sage Pub.

Chelliah, R. K. (2010). A response to the proposal to require registered migration agents to obtain an overall ielts score of 7 in order to obtain re-registration. Received by Michael Suss, Date?

Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension* (Vol. 4). Cambridge University Press.

Coleman, D., Starfield, S., & Hagan, A. (2003). The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australian, .K.U.K. and Chinese tertiary institutions. *IELTS Research Report*, *5*, 160. https://search.informit.org/doi/10.3316/informit.078536418078910

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Report*, *1*, 72-115. https://search.informit.org/doi/10.3316/informit.933870016439796

Davies, A. (1967). The English proficiency of overseas students. *British Journal of Educational Psychology*, *37*(2), 165-174. https://doi.org/10.1111/j.2044-8279.1967.tb01925.x

Davies, A. (1984). Validating three tests of English language proficiency. *LANGUAGE TESTING*, *1*(1), 50-69. https://doi.org/10.1177/026553228400100105

Davies, A. (1999). Standard English: Discordant voices. *World Englishes*, *18*(2), 171-186. https://doi.org/10.1111/1467-971X.00132

Davies, A. (2005). *An introduction to applied linguistics*, (1), 59, 77-79. https://doi.org/10.1093/elt/cci013

Davies, A. (2014). Remembering 1980. *Language Assessment Quarterly*, *11*(2), 129-135. https://doi.org/10.1080/15434303.2014.898642

Douglas, D. (1990). *English language testing in .S.U.S. colleges and universities*. [Report]. National Association for Foreign Student Affairs, Washington, D.C. https://docslib.org/doc/697292/english-language-testing-in-us-colleges-and-universities

Elder, C., Knoch, U., & Kim, H. (2016). *Preparing for the NAATI examination: options and issues for English proficiency screening.* The University of Melbourne.

Ellis, M., Chong, S., & Choy, Z. (2013). IELTS as an indicator of written proficiency levels: A study of student teachers at the National Institute of Education, Singapore. *International Journal of Educational Research*, *60*, 11-18. https://doi.org/10.1016/j.ijer.2013.03.003

Everitt, B. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press.

Fitzner, K. (2007). Reliability and validity: A quick review. *Diabetes Education*, *33*(5), 775-776, 780. https://doi.org/10.1177/0145721707308172

Friedmann, D. (1989). The efficient breach fallacy. *The Journal of Legal Studies*, *18*(1), 1-24. https://www.journals.uchicago.edu/doi/abs/10.1086/468138?journalCode=jls

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *LANGUAGE TESTING*, *26*(1), 123-144. https://doi.org/10.1177/0265532208097339

Gagen, T. (2019). *The Predictive Validity of IELTS Scores: A Meta-Analysis*. (6406) [Master Thesis Dissertation: The University of Ontario] (Electronic Thesis and Dissertation Repository). https://ir.lib.uwo.ca/etd/6406

Green, A. (2019). Restoring perspective on the IELTS test. *ELT Journal*.

Green, T., and Maycock, Louise. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes*, *18*, 3-6. https://doi.org/10.1093/elt/ccz008

Hall, G. (2009). International English language testing: a critical response. *ELT Journal*, *64*(3), 321-328. https://doi.org/10.1093/elt/ccp054

Hamid, M. O. (2015). Policies of global English tests: test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education*, *37*(3), 472-487. https://doi.org/10.1080/01596306.2015.1061978

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 33-41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Hawthorne, L. (2013). *English Language Skills Registration Standards*.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports*, *2*, 62-73. IELTS Australia.

Hirsh, D. (2007). English language, academic support and academic outcomes: A discussion paper. *University of Sydney papers in TESOL*, *2*(2), 193-211. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=edc5ab57e1ec79c4315b22a52b1d5967d81414f7

Hogan, M. J. (2005). *'Something new? Quite a lot in IELTS, actually.'* [Paper presentation]. 18th Annual English Australia (E.A.) Education Conference. www.englishaustralia.com.au/ea_conference05/proceedings/pdf/Hogan.pdf

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkinshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study. *IELTS Research Reports Online Series*, *Series 2012/1*, 41. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2012-1.ashx

IDP: IELTS Australia. (2011). *Submission 155 to the Senate Standing Committee on Health and Ageing Inquiry into the Registration processes and support for overseas trained doctors*. www.aphref.aph.gov.au-house-committee-haa-overseasdoctors-subs-sub155

IELTS. (2004). *Standards against which IELTS must be measured.* http://web.archive.org/web/20040614023009/http://www.ielts.org/teachersandresearchers/standardsandconstructs/article147.aspx

IELTS. (2007a). IELTS - English for International Opportunity.pdf. http://www.ielts.org/teachersandresearchers/analysisoftestdata/article382.aspx

IELTS. (2009). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/pdf/IELTS%20Guide%20for%20Stakeholders%20March%202009.pdf

http://web.archive.org/web/20090501000000*/https://www.ielts.org/pdf/IELTS%20Guide%20for%20Stakeholders%20March%202009.pdf

IELTS. (2011). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/PDF/Guide_Edu-%20Inst_Gov_2013.pdf

IELTS. (2012). *Researchers - Test performance 2012.* http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2012.aspx

IELTS. (2013). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/PDF/Guide_Edu-%20Inst_Gov_2013.pdf

IELTS. (2014a). *Test Performance 2013.* http://web.archive.org/web/20150820122427/http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2013.aspx

http://web.archive.org/web/20151023065504/https://www.ielts.org/researchers/analysis_of_test_data/test_performance_2013.aspx

IELTS. (2014b). *Test taker performance 2014.* https://www.ielts.org/teaching-and-research/test-taker-performance-2014

IELTS. (2015a). *Test Performance 2014.* http://web.archive.org/web/20150912090121/http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2014.aspx

http://web.archive.org/web/20160125074055/http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2014.aspx

IELTS. (2015b). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/PDF/Guide_Edu-%20Inst_Gov_2013.pdf

https://www.ielts.org/~/media/publications/guide-for-institutions/ielts-guide-for-institutions-2015-uk.ashx

IELTS. (2016). *Guide for educational institutions, governments, professional bodies and commercial organisations.* IELTS research reports. https://www.ielts.org/~/media/publications/guide-for-institutions/ielts-guide-for-institutions-2015-uk.ashx

IELTS. (2016a). *Test Performance 2015.* https://www.ielts.org/teaching-and-research/test-performance

IELTS. (2016b). *Institutions - Globally recognised testing system.* http://www.ielts.org/institutions/global_recognition/globally_recognised_testing.aspx

IELTS. (2018). *How Your Score Has Been Calculated?* http://idpielts.me/results/how-your-score-has-been-calculated/

IELTS. (2019a). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-uk.ashx?la=en

IELTS. (2019b). *IELTS scoring in detail.* https://www.ielts.org/ielts-for-organisations/ielts-scoring-in-detail

IELTS. (2019c). *Test Taker Performance 2018.* https://www.ielts.org/teaching-and-research/test-taker-performance

IELTS. (2019d). *IELTS scoring in detail.* https://www.ielts.org/ielts-for-organisations/ielts-scoring-in-detail

IELTS. (2020a). *Guide for educational institutions, governments, professional bodies and commercial organisations.* https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-uk.ashx?la=en

IELTS. (2021a). *Research reports.* IELTS. https://www.ielts.org/for-researchers/research-reports

IELTS. (2021b). *IELTS scoring in detail.* https://www.ielts.org/for-organisations/ielts-scoring-in-detail

IELTS. (2022). *Ensuring quality and fairness in international language testing.* Retrieved 02 February 2022 from https://www.ielts.org/-/media/publications/quality-and-fairness/quality-and-fairness-2015-uk.ashx

Ingram, D. E. (2005). *The use and abuse of IELTS.* http://www.monash.edu.au/lls/China/learning/ingram4.xml

http://web.archive.org/web/20141007084022/http://www.monash.edu.au/lls/China/learning/ingram4.xml

Ingram, D. E., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, Part 1. *IELTS Research Reports*, *7*, 137-204. IELTS Australia and British Council

Johnson, C. (2007). John Howard's 'Values' and Australian Identity. *Australian Journal of Political Science*, *42*(2), 195-209. https://doi.org/10.1080/10361140701319986

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, *3*, 85-108. IELTS Australia.

Kling, H., Nachtnebel, H. P., & Fürst, J. (2007). Hydrological Atlas of Austria–Mean annual areal actual evapotranspiration (3.3). *Lebensministerium (BMLFUW).* https://hess.copernicus.org/preprints/hess-2022-261/hess-2022-261-ATC2.pdf

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic Feedback in Language Assessment. *The handbook of language teaching*, 610. https://doi.org/10.1002/9781444315783.ch32

Lloyd-Jones, G., Neame, C., & Medaney, S. (2012). A multiple case study of the relationship between the indicators of students' English language competence on entry and students' academic progress at an international postgraduate university. *IELTS Research Reports*, 1-54. IDP: IELTS Australia and British Council.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

McNamara, T. (2000). *Language testing*. OUP Oxford.

Mertens, D. M. (2003). Mixed methods and the politics of human research: The transformative-emancipatory perspective. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research (pp. 135-164).* Sage Pub.

Mina, M. (2021). *Understanding the nonverbal communication skills and the four maxims of discourse theories behind the IELTS speaking strategies taught to students and raising students' awareness of the importance of these strategies*, [Master's thesis Dissertation, Quito]. https://repositorio.usfq.edu.ec/handle/23000/11063

Mokhsin, M., Aziz, A. A., Hamidi, S. R., Lokman, A. M., & Halim, H. A. (2015). Impact of using abbreviation and homophone words in social networking amongst Malaysian youth. https://doi.org/10.1166/asl.2016.6674

Müller, A. (2015). The differences in error rate and type between IELTS writing bands and their impact on academic workload. *Higher Education Research & Development*, *34*(6), 1207-1219. https://doi.org/10.1080/07294360.2015.1024627

Nagle, C. L., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Beyond linguistic features: Exploring behavioral and affective correlates of comprehensible second language speech. *Studies in Second Language Acquisition*, *44*(1), 255-270. https://doi.org/10.1017/S0272263121000073

O'Loughlin, K. (2013). *Developing the assessment literacy of IELTS Test users in higher education.* IELTS. http://www.ielts.org/pdf/vol13_report5.pdf

O'Loughlin, K. (2015). 'But isn't IELTS the most trustworthy?': English language assessment for entry into higher education. *International Education and Cultural-Linguistic Experiences of International Students in Australia*, 181.

Palomba, C. A., & Banta, T. W. (1999). *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education. Higher and Adult Education Series*. ERIC.

Pearson, W. S. (2019a). Critical perspectives on the IELTS test. *ELT Journal.* https://doi.org/10.1093/elt/ccz006

Pearson, W. S. (2019b). Remark or retake? A study of candidate performance in IELTS and perceptions towards test failure. *Language Testing in Asia*, *9*(1), 17. https://doi.org/10.1186/s40468-019-0093-8

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227. https://doi.org/10.1126/science.1213847

Perlin, M. L., Harmon, T. R., & Wetzel, S. (2020). Man Is Opposed to Fair Play: An Empirical Analysis of How the Fifth Circuit Has Failed to Take Seriously Atkins v. Virginia. *Wake Forest JL & Pol'y, 11, 451.* https://heinonline.org/HOL/LandingPage?handle=hein.journals/wfjlapo11&div=18&id=&page=

Persaud, N., & Dagher, R. (2021). Evaluation in Our New Normal Environment: Navigating the Challenges with Data Collection. *Journal of MultiDisciplinary Evaluation*, *17*(38), 1-15. https://doi.org/10.56645/jmde.v17i38.673

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafo, M. R., Nichols, T. E., Poline, J. B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci*, *18*(2), 115-126. https://doi.org/10.1038/nrn.2016.167

Punch, K. F. (2013). *Introduction to social research: Quantitative and qualitative approaches*. Sage. https://books.google.com.au/books?hl=en&lr=&id=G2fOAgAAQBAJ&oi=fnd&pg=PP1&ots=j3rKAhaSxo&sig=WY0f117pYuXIofmkwz6QbnmkRE0#v=onepage&q=not%20the%20test%20itself&f=false

Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and Research Integrity. *Account Res*, *24*(2), 116-123. https://doi.org/10.1080/08989621.2016.1257387

Sawand, F. A., Chandio, B. A., Bilal, M., Rasheed, M. R., Raza, M. A., & Ahmad, N. (2015). Quality Assessment in Higher Education. *International Letters of Social and Humanistic Sciences*, *50*, 162-171. https://www.learntechlib.org/p/177133/

Templer, B. (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market. *Journal for Critical Education Policy Studies*, *2*(1), 1-8. https://d1wqtxts1xzle7.cloudfront.net/37941128/templer_2004-libre.pdf?1434668604=&response-content-

Uysal, H. H. (2009). A response to Graham Hall. *ELT Journal*, *64*(3), 329-330. https://doi.org/10.1093/elt/ccp079

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

Watson, R., & Hayter, M. (2020). When nurses ignore evidence. https://doi.org/10.1111/jan.14383

Weir, C. J. (2005). *Language testing and validation.* Palgrave McMillan.

Wilde, S. (2002). *Testing and Standards: A Brief Encyclopedia*. ERIC.

Winefield, A. H., Gillespie, N., Stough, C., Dua, J., Hapuarachchi, J., & Boyd, C. (2003). Occupational stress in Australian university staff: Results from a national survey. *International Journal of Stress Management*, *10*(1), 51-63. https://doi.org/10.1037/1072-5245.10.1.51