“The system is stacked against them”:

Investigating Issues of Fairness in the IELTS Writing Task for Test Takers

Elaheh Zaferanieh

A Thesis in
The Department
of
Education

Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Arts
at Concordia University
Montreal, Quebec, Canada

September 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Elabeh Zaferanieh

Entitled: "The System is stacked against them": Investigating the Issues of Fairness in the IELTS Writing Task for Test Takers

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Educational Technology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair
Dr. Saul Carliner

_____ Examiner
Dr. Steven Shaw

_____ Examiner
Dr. Heike Neumann

_____ Supervisor
Dr. Julie Corrigan

Approved by _____
Dr. Saul Carliner, Chair of Department

_____2024 _____
Pascale Sicotte, Dean of Faculty

Abstract

"The system is stacked against them":
Investigating Issues of Fairness in the IELTS Writing Task for Test Takers

Elaheh Zaferanieh

This research investigated the perceived fairness of the IELTS writing assessment from the perspectives of educators/examiners and test takers. Using a mixed methods approach, the study gathered quantitative data from questionnaires completed by 30 participants and qualitative information from interviews. Quantitative findings revealed dissatisfaction among educators, examiners, and test-takers with the IELTS writing assessment, emphasizing its lack of clarity of assessment criteria and fairness. Qualitative data identified four critical themes for both groups: Unclear Scoring Criteria, Cultural Bias, Life-Changing Consequences, and Technological Impact. Themes showed limitations within the scoring system and their impacts on test takers' futures. Notably, the Life-Changing Consequences theme highlighted deep social impacts of writing scores which delayed university admissions, hindered job opportunities and complicated immigration processes for some candidates. The findings called for a re-evaluation of IELTS scoring criteria and advocated for clearer guidelines to mitigate subjectivity and bias. By tackling these issues, the study sought to improve fairness and lessen the unfair challenges for test takers, aligning the test more accurately with their true language abilities.

Acknowledgements

I owe a debt of gratitude to all people who helped me to accomplish this thesis.

Thanks are due to my dear supervisor Dr. Julie Corrigan for her invaluable help and constructive suggestions.

I would like to thank my committee members Dr. Steven Shaw and Dr. Heike Neumann for their special contributions and helpful comments.

Finally, I wish to thank my family for their support and patience.

Dedication

To my dear mother, for all her love and support,

and

to Dr. Julie Corrigan, who has been more than a supervisor by helping me

in every aspect of my life

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

For those people who intend to immigrate, work or pursue their education in English speaking countries, the International English Language Testing System or IELTS is a very important test, which is globally recognized and provides access to opportunities for education and immigration. Like other high-stakes exams, it has had a big impact on people's social, economic, and educational lives since outcomes of these tests are typically used to make important decisions about things like hiring, immigration and university admission (McNamara, 2005).

According to the IELTS website (2024), IELTS scores serve the following purposes: they assess English communication skills for academic, professional, and immigration needs. These scores are accepted by universities, professional bodies, and governments for admissions, accreditation, and residency applications, particularly in countries like Australia, New Zealand, Canada, the UK, and the USA.

Among all the sections in an IELTS test, the writing test is particularly notable for having a significant effect on test takers' final results and in turn, their prospects (Hamid *et al*., 2019). This section of the IELTS assesses a person's capacity to write in English clearly and logically, to cover variety of topics from vocabulary and syntax to concept structure and reasoning, and to address different tasks and various assessment criteria.

Due to the real and important consequences stemming from the use and interpretation of IELTS scores, it is crucially important that such a test meet the highest standards of validity and reliability according to the *Standards for Educational and Psychological Testing* (henceforth, the *Standards;* AERA/NCME/APA, 2014*)* but also of fairness. A full consideration of the topic would explore the multiple functions of testing in relation to its many goals, including the broad goal of

achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways in which test results are reported and used, the factors that affect the validity of score interpretations, and the consequences of test use. (p. 49). This definition establishes the principle of fair and equitable treatment of all test takers during the testing process. The second, third, and fourth views presented here emphasize issues of fairness in measurement quality: fairness as the lack or absence of measurement bias, fairness as access to the constructs measured, and fairness as validity of individual test score interpretations for the intended use(s) (pp. 5-51).

Fairness is interpreted as responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses (NCME, 2014). This means that all applicants must be examined equally, with their language and writing skills taking precedence over other variables such as the test taker's background, test-taking location, or the specific examiner's personal biases. In this case, fairness is not just necessary technically, but morally essential as well (Kunnan, 2000; Xi, 2010).

Test fairness is a particular concern for academics, test, and educators. In *The Standards*, and according to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014), fairness is considered a crucial test quality and involves treating all test takers equally. Fairness in treatment during the testing process, fairness as a lack of measurement bias, access to the construct(s) as measured, and fairness as validity of test score interpretations for the intended uses, are also taken into consideration. Numerous scholars, such as Fulcher and Davidson (2007), Weir (2005), and Shohamy (2001a, 2001b), believe that exams should be as fair as possible and that their fairness should be evaluated, particularly when they are high-stakes exams with substantial ramifications (Stobart, 2005). Considering the important decisions and significant life

choices based on high-stakes tests, research is needed to ensure the fairness of these tests. The consequences of a high-stakes test may become more serious when it comes to international tests and test-takers. In this regard, researchers such as Slomp et al. (2014) have put emphasis on consequential aspects of large-scale writing tests by proposing a model of consequential validity research.

Fairness in the IELTS writing exam can be related to construct-irrelevant barriers, such as the prompts' consistency and clarity, the impartiality of scoring criteria, and the avoidance of language or cultural biases. While these factors are critical for ensuring fairness, they also impact the reliability of the test. According to Rudner and Schafer (2002), clarity of test items and consistency in scoring reduce measurement error, thereby improving the reliability of the results. A reliable test is essential for ensuring that results accurately reflect a test-taker's true abilities. At the same time, fairness is achieved when a writing assessment is impartial and free of prejudices that might advantage or disadvantage particular groups, thus accurately reflecting each test-taker's proficiency in writing in the English language (Uysal, 2010).

Background

Test Fairness. As fairness is a complex concept, so far, it has been conceptualised and defined in various ways, and despite its fuzzy definitions, some researchers have tried to take concrete approaches to test fairness.

In *The Standards for Educational and Psychological Testing*, measurement techniques are discussed in detail. Regarding fairness in educational assessment, in the first edition of this book (1985), the concept of fairness was not defined. In the second edition (1999), fairness was defined

as “the principle that every test taker should be assessed in an equitable way” (p. 175), and in the last edition in 2014, fairness was regarded as “The validity of test score interpretations for intended use(s) for an individual from all relevant subgroups. A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals” (p. 219). In this edition, fairness was given approximately equal standing as reliability and validity, and a separate chapter was allocated to it.

For conceptualizing and evaluating fairness, some researchers focus on the relationships between validity and fairness as a concrete approach. For instance, Xi (2010) regards fairness as a part of a validity argument. She mentions that this relation is usually conceptualized in three ways: fairness independent from validity, fairness including validity, and fairness a major part of validity.

Among the definitions offered so far, one of the most influential ones, derived from techniques in *The Standards for Educational and Psychological Testing*, was proposed by Kunnan (2000). He introduced a framework which was based on ethics and involved three test qualities, namely validity, absence of bias, and social consequences. However, the framework was not complete as it lacked accessibility and administration considerations. Thus, Kunnan proposed a new version of the fairness framework in 2004. The most recent ones (Kunnan, 2004, 2010) had two more factors, namely access and administration. In this comprehensive framework, fairness is regarded as a whole system that is beyond a test. It includes many sides, ranging from test use (both intentional and unintentional consequences and use) to test development steps and even different stakeholders involved in the test process.

After these definitions and models were introduced, in 2011, McNamara and Ryan, referring to Messick’s (1989) seminal theory, proposed a distinction between fairness and justice.

They believed that fairness is a technical and internal quality in language assessment, with an evidential basis, whereas justice is an external quality with a consequential basis. They argued that approaches such as those by Xi's (2010) do not address the consequential aspects separately, while direct arguments on consequential dimensions are necessary.

Considering these distinctions and definitions, Kunnan's (2004, 2010) fairness framework seemed to be multi-faceted, and by having detailed specifications for each facet, it became an appropriate model for researchers who studied test fairness.

Research on fairness. Pishghadam and Tabataba'ian (2011) examined fairness by considering the relationship between IQ and course test format. This study focused on whether different test formats favor individuals with varying IQ levels, examining how IQ impacts test performance across multiple formats. Rather than evaluating the inherent fairness of the test itself, the research addressed differential performance—exploring if certain formats advantage or disadvantage individuals based on cognitive abilities. This approach relates to fairness in test outcomes, rather than assessing bias in the test's design or content.

In more recent research by Hamid et al. (2019), the researchers tried to examine the fairness issues in International English Language Testing System (IELTS) from test takers' perspective. It was found that test-takers perceive significant issues of fairness, justice, and validity, influencing their views on the test's impacts. This study was a pioneering study in investigating fairness in language proficiency tests. In spite of the fact that they focused on fairness, their investigation was limited to socio-political aspects of fairness.

So, a standardized test, like the IELTS, may actually favor certain linguistic structures or cultural knowledge bases with which some candidates are unfamiliar. For example, test prompts referencing Western holidays or traditions may disadvantage candidates from non-Western

backgrounds. Similarly, some candidates may struggle with essay structures more common in Western academia, such as explicit thesis statements. Fairness in the test will thus be affected when some demographic groups perform better simply because test design is subtly more aligned with the educational norms of certain linguistic or cultural groups. Such biases can lower the fairness of the test as a measure of true language proficiency.

Furthermore, the role of examiners in scoring writing tasks adds another layer of complexity into considerations of fairness. As Weir (2005) notes, examiner subjectivity can result very strongly in the scores when the test requires a judgment which is subjective in nature, as in essay writing. The standardization of the criteria and rigorous training of the examiners become crucial in reducing this subjectivity. The detailed rubric of IELTS and frequent re-calibration sessions by examiners work to re-align their judgments with the standard benchmarks, thereby trying to eliminate individual biases in scores.

Finally, fairness could be affected by the availability of test preparation resources. Green (2007) has discussed in a paper how this kind of access to coaching or specialized preparatory materials might actually enhance candidates' scores greatly in certain sections, such as writing. This will result in an unlevel playing field where those with more resources may score higher. This is not because of language ability per se, but because of increased test preparedness. Addressing these disparities requires ongoing adjustments to the test's design and access to preparation resources to ensure that all potential test-takers have a fair chance of succeeding based on their true proficiency.

These detailed discussions thus underpin the urgent need for continued research into the procedures for IELTS writing assessment, if these practices are to stay fair and valid. Everything

from designing test prompts to training examiners and test preparation resources must be subject to regular review and updated on the basis of research and fairness audits.

Although a substantial body of research has been dedicated to exploring the fairness of language assessments, a notable gap persists in the literature regarding the subjective aspects of fairness. Most studies to date have predominantly concentrated on quantifiable elements such as score reliability, predictive validity, and statistical bias, employing quantitative methods to offer empirical evidence of test fairness. These approaches while valuable for their precision and objectivity, often overlook the nuanced, lived experiences of test-takers and the perceptual dimensions of fairness that qualitative methods are well-suited to capture. So, research with qualitative methodologies is needed to elucidate the subjective interpretations of fairness among various stakeholders in the assessment process (Hamid *et al.,* 2019).

In a systematic review study which was conducted to analyze the methodological approaches related to fairness issues in language assessment research, it was found that a predominant number of studies have utilized quantitative methods, and specifically, among the studies that investigated the writing construct, a mere 29.17% used qualitative research methods (Zaferanieh, 2023). The dominance of quantitative research in this field underlines the findings discussed earlier where the focus has been mostly on objective measures like score reliability and predictive validity. The limited use of qualitative methods further highlights the gap which exist in exploring the subjective experiences and perceptions of test-takers. These figures back the argument for a greater incorporation of qualitative research designs, which could provide a deeper understanding of fairness from the perspective of those directly affected by language assessments. This evidence strengthens the case for a paradigm shift towards embracing the complexity and subjectivity inherent to the concept of fairness.

The Role of Technology in IELTS Writing

The use of technology in language testing in general, and in the IELTS writing component in particular, is a paradigmatic shift in how tests are conducted and scored. Test delivery and scoring through digital platforms will add increased accuracy and consistency, two elements cardinal to the reliability of a high-stakes test such as IELTS. Technology in testing may bring about substantial improvements in the process of administration, which then becomes more efficient and less liable to human error.

However, with the technology also comes a set of different issues, that have to be very carefully managed for fairness to be maintained in the test. Of important concern is the digital divide. This might also entail that access to technology is unequal, thus potentially skewing the results toward those who are more adept technologically or better prepared for the examination with expensive tools. This aspect is very important within the context of IELTS, as this deals with educational and immigration opportunities.

The role of technology in IELTS writing tests has therefore to be viewed against multiple parameters. By addressing these, one would want to ensure that the technological integration into IELTS works to further educational equity, rather than inadvertently erecting barriers for certain sets or groups of test-takers.

In summary, while technology affords some opportunities for improving the logistical aspects of test administration and increasing scoring objectivity, it must also raise a critical review of its broader implications for fairness and accessibility. That all candidates have an equal opportunity to do well in the writing test of IELTS, regardless of ability or access to using digital

technologies, remains paramount. This balance will be important in ensuring the integrity and fairness of the test in technologically changing times.

Statement of the Problem

This study is motivated by a combination of research gaps, urgent concerns over the fairness of high-stakes language exams, and the changing role of technology in testing procedures. Even though the IELTS writing test is crucial in shaping the lives of a great number of people globally, a number of important problems have arisen that show the need for this study.

Inadequate Knowledge of Fairness Perceptions. The substantial amount of research on the subject offers little information about how teachers and examiners view fairness in the IELTS writing test. These gaps are important because stakeholders' perspectives have direct impacts on a test's design, administration and scoring. To find any possible discrepancies in how fairness is operationalized, which can result in inconsistencies in test administration and scoring, a more thorough investigation of their perspectives is essential (Hamid et al., 2019).

Underexplored Difficulties for Non-native English Speakers. While earlier studies mentioned the general difficulties non-native English speakers encounter in language exams, in-depth studies that concentrate especially on the IELTS writing section are very few. To my knowledge, there are very few empirical studies that focus on fairness in the IELTS test, and even those that exist have not employed mixed methods as a comprehensive approach. Considering the wide range of backgrounds and particular difficulties that IELTS test-takers encounter when

attempting to demonstrate their writing skills, this gap is especially troubling (Arefsadr & Babaii, 2023).

Emerging Impact of Educational Technology. Although there has been a significant evolution in use of educational technology in language assessment procedures, little is known about how this will affect the fairness of the IELTS writing score. It is important to comprehend how digital tools and platforms such as computer-based Academic IELTS test affect the fairness of language assessment procedures as they become more prevalent in the delivery and the scoring of these evaluations. It is important to evaluate if technology developments are contributing to the improvement of scoring methods' objectivity and equity or whether they are bringing in new types of bias (Azizi, 2022).

This work has been inspired by the gaps in research literature and the important questions they raise regarding the fairness of the IELTS writing assessment. Research attempts to contribute to more thorough knowledge of fairness in high-stakes language testing by addressing these particular issues.

Significance of Study

This study is significant for a number of reasons, all of which highlight how much it can add to the domains of language assessment, educational technology, and policymaking. This research aims to fill current gap in literature about subjective experiences and views of test fairness among educators, examiners and test-takers. This knowledge is essential for creating testing procedures that are inclusive of technology like computer-based exams, cultural sensitivity and language accuracy.

Furthermore, by examining the particular difficulties encountered by test takers in IELTS writing element, the study is expected to provide insights that may result in more inclusive and fair test designs. Given the IELTS's worldwide reach (British Council, n.d.) and its influence on the academic and professional careers of people from a variety of language and cultural backgrounds, this study's component is very important.

Additionally, this study addresses a developing field of interest in language evaluation by looking at how educational technology as used in computer-based exams affects the fairness of IELTS scoring procedures. The results should guide the use of technology in language testing, making sure that any developments in this area promote objectivity and fairness rather than create new sources of prejudice.

Research Objectives

Through three main goals, this study seeks to evaluate fairness in IELTS writing tests. It aims to evaluate educators' and examiners' perceptions of the fairness of IELTS writing test. Second, the study aims to describe and explore the particular difficulties encountered by test takers. Finally, the study will investigate how educational technology through computer-based exams affects scoring equity and consider whether or not technological developments strengthen the evaluation of writing abilities. These goals work together to direct a comprehensive inquiry into enhancing fairness in high-stakes language testing.

Research Questions

This study will address the following questions:

1. How do educators and examiners perceive the fairness of the IELTS writing assessment criteria?

2. What challenges do test takers face in IELTS writing assessments in terms of fairness?

3. How do technological tools and resources influence the fairness of scoring in IELTS writing assessments?

Operational Definitions

To avoid ambiguity and to enhance precision in the study the following operational definitions are put forward for key words and concepts used in this research.

***IELTS Writing Assessment*:** Refers to writing component of International English Language Testing System, specifically Task 2. This task assesses test-takers' ability to write well-structured argumentative essay expressing complex ideas clearly in written English.

Fairness in Testing: Fairness in this study is defined as extent to which IELTS writing test offers equal opportunities for all candidates for showcasing their true language ability without bias. This involves ensuring that scoring criteria are transparent, consistently applied and free from any prejudice toward test-takers' linguistic, cultural or educational backgrounds. Additionally, fairness will consider how technology impacts the equity of these evaluations (Kunnan, 2004).

Bias: A form of error in testing that unfairly advantages or disadvantages specific groups of test-takers. Bias in this study could arise from cultural assumptions in the test prompts, examiner subjectivity or unequal access to test-preparation resources.

Perceived Fairness*:* Subjective judgment of fairness made by test-takers, educators and examiners regarding IELTS writing assessment. This includes perceptions of transparency of

scoring criteria, consistency in scores and presence of biases in test content or processes (Hamid *et al.,* 2019).

Limitations

This study was designed with some limitations that define scope and focus of research. First, study concentrated specifically on IELTS writing assessment rather than addressing all sections of IELTS exam. This decision was made to allow for a more in-depth exploration of perceived fairness and cultural biases in writing section, which has been identified as a key area of concern in previous research. Additionally, this study focused on adult test-takers who had previously taken IELTS exam excluding participants who had no prior experience with test. This research also relied on self-reported data from test-takers and educators without doing a direct analysis of actual IELTS essays or examiner feedback. Also, interviews and questionnaires were conducted in English which may have influenced participants' ability to express opinions. Finally, while this study aimed for a geographically diverse sample, it primarily recruited participants from a limited number of non-native English-speaking countries, which may limit the ability to generalize the findings to other non-native English-speaking regions that were not well-represented in the sample. While these aspects were intended to narrow the study's focus, the main limitation is that this research considered perception data in isolation, without integrating it with other variables (e.g., bias analysis or Rasch analysis of rater scores, or analysis of training materials). The lack of these additional analytical methods may limit the ability to draw broader conclusions about the systemic sources of bias or inconsistencies in examiner scoring.

The Thesis Structure

This thesis consists of five chapters each of which serves a particular purpose in examination of the IELTS writing tests for fairness. Chapter 1 introduces study's background, problems, significance, objectives and research questions. It also sets the central operational terms in use. Chapter 2 provides an in-depth review of literature from studies relating to concepts of fairness and bias in language assessment, previous research into IELTS and other major high-stakes tests, and the role of technology in language assessment, with particular regard to Academic Task 2. Chapter 3 outlines the methodology that was employed in which the research design, participant selection, data collection methods, and approaches taken toward the analysis of data are explained. The ethical considerations of the study are also shed light on. Chapter 4 presents the findings where the analysis for both data from interviews and data from questionnaires is done and integrated into a wholistic understanding of these issues. Then, the results are interpreted in view of the prevailing literature, discussed at the theoretical level. In chapter 5, conclusions, implications of the results and practical suggestions towards test design and policy are presented. It also discusses the limitations of the study and suggests some future research directions. The appendices that follow include the interview guide, questionnaire, and consent forms used in the research study.

Conclusion

The goal of this study is to critically examine issue of fairness in IELTS writing tests which is an important component for anybody looking for opportunities abroad. It seeks to address issues for non-native English speakers, discover the complex views of fairness held by many stakeholders, and investigate how technology affects assessment procedures. The study, which takes an in-depth qualitative approach, attempts to fill the gaps in the field by providing insights

that have a big impact on testing procedures, educational policy, and the use of technology in language assessments. This study looks to offer helpful ideas on how to ensure fairness in language assessments with significant consequences, with the goal of creating a more inclusive and a more equitable evaluation process for all applicants.

Chapter 2

Review of Literature

Introduction

The test of language proficiency stands as an inevitable part of the global educational scenario, particularly in admission, employment, or immigration to an English-speaking country. In the midst of all such tests for language proficiency, the International English Language Testing System holds a marked and recognized place among the most noted and duly taken tests, which opens doors of opportunities and guides the very choices of educational and professional decisions of millions of test-takers worldwide (British Council, 2023; Green, 2007).

The IELTS test which has two modules, academic and general, targets a language skill set: listening, reading, speaking, and writing. In both of these modules, for writing skill, there are two tasks among which task 2 is of paramount importance. This is because it not only takes up a more significant mark component than the others but is also very instrumental in both academic and professional achievement (Moore, Morton, & Price, 2015). This section requires test-takers to write a well-structured, argumentative essay that demonstrates the ability to formulate complex ideas and express them in written English. Given the high stakes of the candidate's performance on IELTS writing, fairness should come first.

Fairness in language testing is very elusive and multi-dimensional. It covers clarity of the scoring criteria, uniformity of the scoring procedure, and freedom from bias stemming either from linguistic, cultural, or educational background (Kunnan, 2010; Xi, 2010). Fairness is not just a technical requirement; it is also an ethical mandate. This is because it tries to protect all test-takers

from conditions that would deny them the opportunity for a fair display of their actual language proficiency (Hamp-Lyons, 2001).

While fairness in the IELTS writing assessment is one of the most critical issues, it has so far been sparsely researched from various educator and examiner perspectives, and especially from that of test-takers. Previous research has primarily focused on fairness in the quantitative sense, with emphasis on psychometric measures such as score reliability and predictive validity, which are central to validity studies in language assessment. These studies often examine the consistency and accuracy of test scores, ensuring they reflect the intended abilities and can predict future performance. However, they have given little attention to the subjective experiences and perceptions of the people involved in the process. This gap in the literature further justifies a more comprehensive exploration of fairness in the IELTS writing assessment, integrating both qualitative and quantitative approaches, to understand how fairness is perceived by test-takers, educators, and examiners beyond traditional psychometric frameworks.

It is another dimension of complexity added to the issue of fairness that technology is being infused into language testing. On one hand, technological innovations like computer-based tests bring in the promise of consistency and efficiency; on the other hand, they also open up new challenges and possible biases. The digital divide issues are concerns that are important to be addressed in the quest for continued fairness in the IELTS writing assessment.

This literature review, conducted for the research, aims at an in-depth review of available findings on fairness in the writing sections of the IELTS test, especially Task 2. Precisely, this chapter seeks to locate the current study within the context of language testing research through discussion of certain concepts and views on fairness and bias, empirical studies conducted on IELTS, other similar tests, and the role of technology in language assessment. It is, therefore, a

foundation of the ensuing chapters for this thesis to review, which will then go on to cover an empirical investigation into the fairness of the IELTS writing assessment.

The structure of this literature review begins with an overview of the IELTS test, focusing on its significance in both academic and professional settings, with particular attention to Task 2 in the writing section. The review then delves into key concepts of fairness and bias in language testing, offering definitions and discussing how these concepts influence design and implementation of high-stakes tests like IELTS. Following this, the chapter reviews relevant empirical studies that investigate fairness of the IELTS writing test, discussing issues such as cultural bias and examiner subjectivity. The discussion then shifts to the role of technology in language assessments, exploring benefits and challenges associated with computer-based testing. The theoretical framework, which includes Kunnan's Test Fairness Framework and other relevant theories, is introduced next, providing a foundation for analyzing fairness in the IELTS writing assessment. Finally, the chapter concludes with a summary of the key findings and sets the stage for the empirical investigation to follow.

What Is IELTS?

The International English Language Testing System (IELTS), one of the most popular tests, is taken to show one's proficiency in English. It therefore plays a crucial role in determining the amount and type of educational, professional and immigration-related opportunities available to persons who wish to go to English-speaking countries either to study, to work, or to settle. The IELTS is jointly managed and delivered by the British Council, IDP: IELTS Australia, and Cambridge Assessment English. The ability to understand and use forms and patterns of the English language is tested in four areas: listening, reading, writing, and speaking. The test comes

in two modules: the Academic module which is intended for those people who are searching for admission either to a higher education institution or to professional registration, and the General Training module, intended for those who want to migrate to an English-speaking country or undertake non-academic training or work experience (British Council, n.d.).

Focus on IELTS Writing Task 2

Writing Task 2 is an essay writing component of the IELTS writing test that holds much value in academic and professional circles. This part of the IELTS is a task that is designed to examine the candidate's ability to put forward and defend an argument, present complicated ideas, express a clear, coherent and cohesive writing style on a topic or as a response to a given prompt. Typically, candidates are required to write an essay of at least 250 words within 40 minutes on a given topic that presents an issue from more than one point of view, gives solutions to a problem or offers a balanced argument. The essay is evaluated for four prime criteria: Task Response, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy—which all add up to the final band score (British Council, n.d.).

One of the major skills in writing is that one should be able to construct a well-organized and convincing essay. Writing Task 2 has been specifically developed to reflect those sorts of writing tasks which people come across in their work settings or their higher education, for example, writing research papers, essays, and reports. In many universities and colleges, high bands in the IELTS writing section, particularly Task 2, remain a key requirement for admission. According to Barkaoui (2016), Task 2 involves the assessment of features such as fluency, syntactic complexity, coherence, and register. These skills are essential for effectively structuring arguments and expressing complex ideas, not only in academic contexts but also in professional environments where clarity and formal writing are critical.

If candidates want to develop in their career, they will find that a strong performance on the writing Task 2 will be very beneficial in professional circles. IELTS scores are one of the benchmarks used by employers and several professional registration organizations in the majority of English-speaking nations to evaluate the language competency of prospective workers (IELTS, n.d.). A high score in this section and more so in Task 2 of the writing would, therefore, indicate the candidate's ability to effectively communicate ideas in written English—the most critical language skill in fields such as law, business, engineering, and medicine. According to Green (2007), "The writing tasks are designed to reflect what professionals may have to do in the real world: write a draft report, make recommendations, or argue a case in writing."

Concepts of Fairness and Bias in Language Testing

Definitions and Theories of Fairness

Fairness in language testing has taken different conceptualizations, thus showing its multifaceted nature and criticality in ensuring that outcomes are fair to all test takers. If one wants to make sure that the results are fair to all candidates, fairness should be dealt with at the development and administration stages of tests supposed to not only be valid but also fair and inclusive.

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) use the term fairness in relation with reducing construct-irrelevant variance. This means that there may be things affecting test scores that are irrelevant to what is measured. The standards suggest that fairness embodies the idea of test scores having valid interpretations for all examinees regardless of subgroup membership. It simply shows the efforts of getting rid of all other variables that may influence test scores.

Xi (2010) integrated the views of many about fairness, proposing that fairness should be evaluated with respect to validity. In this respect, validity refers to the extent to which theory and evidence support interpretations of test scores for intended purposes. Xi puts fairness as a component of validity, for a test should not be valid if it is not fair to one or more groups of test-takers. Such a view concurs with the views of Messick (1989), in that validity relates to all aspects of a test, such as fairness and the consequences which may come as a result of test use. Xi (2010) categorizes the strategies to integrate fairness and validity into three: fairness independent from validity, fairness including validity, and fairness as major part of validity. On one hand, fairness has been considered to be entirely separate from validity. On the other hand, fairness could be conceived as a component of validity. Lastly, from a more comprehensive view, fairness is described as part of validity as it significantly contributes to assessing the overall effectiveness and appropriateness of the test.

Along these lines, a variety of approaches at the theoretical level have been worked out for a further understanding and assessment of fairness in language testing. One such influential model is Kunnan's (2004, 2010) extended fairness framework: access, administration, validity, absence of bias, and social consequences. It is a framework that provides an overview of fairness in view of factors like access to the test by any person, procedures of fairness in test administration, the validity of the test, and lack of bias in test content and scoring. It also looks at the social consequences of the test, noting that tests have major impacts on people and society.

Kunnan in 2000 suggested an ethical explanation for test fairness framework (TFF) which included some principles and sub-principles. The framework had three test components which were validity, absence of bias, and social consequences. Then, in 2004, Kunnan proposed a new framework in which components of access and administration were also added.

In this framework, Kunnan (2004) viewed fairness as a whole system of a testing. Therefore, many facets of fairness were at stake such as tests uses for planned and unplanned purposes, many stakeholders in the process of testing such as those who take the test, test users, educators, and employers, and finally many stages in the process of developing the test like test design, development, administration, and use.

Underlying this farmwork, there were also two main principles: justice and beneficence. The notion of justice tries to guarantee that a test is fair to all people who take the test. This principle itself has two sub-principles: first, any test should have comparable construct validity in regard to test-score interpretation for all test takers; second, it should not be biased against any test taker categories, particularly by examining construct-irrelevant factors.

Moreover, according to the notion of beneficence, a test should result in positive social outcomes. In other words, it should not be dangerous or destructive to society, and it has to advance beneficial developments in society by delivering test score information and social advantages. It should not cause harm by giving misleading or inaccurate details about exam scores or social consequences.

Considering the mentioned principles, test fairness framework has five main components, including validity, absence of bias, access, administration, and social consequences *(*Kunnan, 2004).

In this framework, validity is one of the important qualities and can be explained as content representativeness, construct validity, criterion validity, and reliability.

The second element which is absence of bias is related to differential consequences, standard setting, and test language or contents. First and foremost, this necessitates changing any test language, content or dialect that offends or discriminates against test-takers from

different backgrounds (due to factors such as gender, race and ethnicity, religion, age, native language, national origin, or sexual preference. Second, differential item functioning (DIF) or differential test functioning (DTF) should be analyzed to determine whether test-takers from different group memberships perform differently due to factors unrelated to the construct being measured. The goal is to ensure that test-takers from various groups do not have distinct performances or outcomes caused by bias. Absence of bias in this context means that performance differences, if they exist, are linked solely to the construct being tested and not to external factors such as gender, race, or language (Kunnan, 2004). Regarding standard setting, test results should be analyzed in relation to the criterion measure and selection decisions. It is essential for test developers and test users to have confidence in the proper metrics and unbiased, reliable selection models are being employed. Test developers and users of scores should be able to deduce from these studies that group differences are linked to the abilities being tested rather than factors unrelated to the construct.

Considering access as the third element, those who develop the tests should ensure that there is access to conditions or equipment as well as educational, financial, geographical, and other resources. This means that testing should be accessible to test takers financially, geographically and in terms of their ability to learn the subject matter. It should also provide certified test takers with physical and learning disabilities with the necessary test accommodations, and it ought to make sure that test takers are familiar with the tools, processes, and conditions used during the test.

The next TFF module, administration, can be described as uniformity in test administration by observing consistency across test locations; equivalent forms and instructions; proper test security; and appropriate physical conditions, such as perfect light, temperature, and facilities.

Finally, social consequences are the last element of TTF. Test developers have to think about how a test would affect and attempt to undo any negative effects by rescoring and reevaluating test responses as remedies or revising test prompts or providing additional accommodations (Kunnan, 2004).

Distinction Between Fairness and Justice in Language Testing

McNamara and Ryan (2011) borrow from Messick's seminal work to posit the distinction between fairness and justice in language testing. In this conceptualization, fairness is internal and technical, whereas justice is considered an external, social, and ethical concern. This work's focus thus resonates with this multifaceted nature of fairness in testing, necessarily going beyond technical issues of validity to address significant social consequences and ethical imperatives stemming from test use.

Fairness in the narrow sense has to do with the internal features of test design and implementation. This relates to making sure that the test measures what it purports to measure without bias, that scoring procedures are consistent, and objective. According to Messick (1989), fairness is basically about validity—the degree to which evidence and theory support the interpretations of test scores for their intended purposes. This internal focus toward evidence and evaluation is aimed to make test scores really be a true reflection of the test taker's abilities and not be contaminated by irrelevant factors; for instance, cultural or linguistically based. For instance, Kunnan (2000) emphasizes that fairness is not bias and encompasses social consequences as part of the validation program. The *Standards for Educational and Psychological Testing* also comment that fairness includes minimizing the construct-irrelevant variance—differences between

the observed test scores that have nothing to do with the construct being measured but are a function of other extraneous factors, such as socio-economic status or educational background.

Justice, on the other hand, refers to other outward implications of the test. This refers to how one's testing impacts them as well as society in terms of social equity and ethical use. McNamara and Ryan (2011) were of the view that if the issue of fairness mostly concerns the technical accuracy and impartiality of a test, then justice looks at the wider consequences of testing practices. This refers to who would be advantaged from the test and who would, on the other hand, be disadvantaged by the same.

The concept of justice in testing is clearly aligned with the concept of consequential validity, as was put forth by Messick (1989). The consequentially valid procedure is hence that which assesses the broader effects of test use, including social and ethical implications. E.g. as Elana Shohamy (2001b) describes the ethical dimensions related to high-stake tests, stating that one has to ponder over how tests can reinforce inequalities or lead to social stratification.

Incorporating both fairness and justice in language testing requires an approach to cover the technical and social dimensions of testing. This will necessarily involve the establishment of validity and reliability of the tests, while staying sensitive to their broader social consequences. Bachman and Palmer (2010) argue in this regard that an ethical approach to language assessment should be based on factors of fairness and justice. The authors view test-making and test-administering personnel as having an ethical responsibility to consider the larger implications - such as how they contribute to testing subjects' lives and opportunities - of the tests they develop and give.

Furthermore, Spolsky (1995) highlights the ethical responsibility of test developers, that their tests should not operate in a manner that disadvantages unjustly any one group. Additionally,

the problems would not concern only the technical questions of bias and validity, but also how the test was more widely used and its broader social consequences. In fact, such an integrated approach is necessary to ensure that language tests are not only technically sound but socially just.

Bias in Language Testing

Bias is a very integral part of fairness in consideration of the fact that it represents the mistake in the testing process that consistently favors or does not favor some specific group of test-takers. Bias can come from many sources to include cultural assumptions attached to the test prompts, subjectivity on the part of the examiner, and access to test-preparatory materials by the test-takers (Shohamy, 2001a, 2001b). Bias has to be reduced to allow every test-taker to operate in an environment in which they have a fair chance to clearly present their ability to have clear and effective language use.

Cultural Bias. Cultural bias occurs when test content reflects differentially the cultural norms, knowledge, and values of one group over another and, hence, tends to prejudice those from other cultural backgrounds. For example, a writing topic related to Western educational methods would be relatively easy to understand and respond to by candidates from Western countries, while the same would be very difficult for candidates coming from a non-Western background. There may be some degree of bias contained in the standardized tests, and the IELTS is an example, with a leaning in favor of some linguistic structures or cultural knowledge bases that might not be familiar to all the test-takers. Hence, such biases can be said to undermine the test's reliability as a true language proficiency measure.

Examiner Subjectivity. Examiner subjectivity plays some role in scoring a writing task because these kinds of scoring depend much on the application of subjective judgments in the precising of the scores of the tests (Weir, 2005). The unconscious bias of the examiners or the way they interpret an item may guide examiner subjectivity. The scoring criteria, therefore, must be standardized, and the examiners must have rigorous training. Taylor (2009), while discussing this topic, argues that the adoption of highly specific scoring rubrics and reconceptualization sessions ensures that examiner alignment against standard benchmarks is done in a regular and periodic fashion. This kind of approach reduces individual biases in scores and results in much more consistent and fairer scoring.

Access to Test Preparation Resources. Disparities in the test preparation environment also contribute to bias. It can make it an unevenly balanced field in which candidates with more resources have the potential to perform better; this would not be based solely on their ability to use the language but partly due to how prepared for the test they were. It involves endless fine-tuning in the design of the test and also access to preparatory materials to put all future test candidates in an equal position to have opportunities to succeed based on their actual proficiency (Green, 2007).

Bias in language testing can lead to significant effects of test outcomes, hence possibly including unfair advantages or disadvantages to one group of examinees over another. A useful approach in the detection of item bias is differential item functioning, which determines whether the items used in measurement function differently on the measurement of different categories of examinees (Camilli & Shepard, 1994). For instance, an item that requires using particular cultural knowledge might be a disadvantage for test-takers with diverse cultural backgrounds. Very many tests have shown that the non-native speakers clearly get lower scores on culturally biased items, which affect their overall test scores and subsequence opportunities.

While bias may not always be immediately evident to individual test-takers, its broader impact becomes apparent in the educational and professional opportunities that may either be granted or denied as a result of biased testing practices. This can result in the misclassification of a candidate's abilities, leading to unjust outcomes, such as denying admission to a qualified student or inaccurately assessing the language proficiency of a job applicant. Consequently, ensuring that language tests are free from bias through rigorous design and validation processes is essential for maintaining fairness and providing equitable access to opportunities.

Previous studies on Language assessments and fairness

Fairness in language tests, including IELTS, has been a primary concern in several studies, most of which are non-empirical, particularly in relation to the productive skills of writing and speaking. These studies have primarily focused on identifying biases and inconsistencies in the test instruments.

Among the influential works on fairness in language testing is Shohamy's *The Power of Tests* (2001b), where she critically addresses the biases present in high-stakes language tests. Her work focuses on how tests can disadvantage test-takers from diverse cultural and linguistic backgrounds due to the inherent cultural biases in test design. Shohamy highlights the importance of creating more inclusive tests that account for cultural diversity to ensure fairness in language assessments.

Shohamy's research emphasizes how cultural factors can affect test performance and the overall fairness of language testing. She argues that many language tests, particularly those used in high-stakes contexts, incorporate culturally specific content related to English-speaking countries. This

content, such as references to Western holidays, customs, or historical events, may be unfamiliar to non-native speakers from other cultural backgrounds. As a result, test-takers may struggle to comprehend the tasks, leading to lower scores that do not accurately reflect their true language abilities.

Shohamy also critiques how these biases are especially pronounced in the writing and speaking sections of language tests, where candidates may be required to respond to culturally loaded topics. It seems that unfamiliarity with such topics may increase test anxiety and decrease confidence, which in turn negatively impacts performance. Such issues can ultimately harm the credibility and acceptance of language tests as fair tools for assessment, particularly when they are used to make critical decisions about education and employment.

Similarly, in another work, in the article "Democratic Assessment as an Alternative", Shohamy (2001a) critiques the use of language tests as powerful tools that have far-reaching consequences on individuals and educational systems. She argues that tests, often used as instruments of power by authorities, are introduced in ways that manipulate educational outcomes and impose specific agendas. These uses of tests can be undemocratic and violate key principles of fairness and inclusion, particularly in multicultural societies. She proposes democratic assessment strategies aimed at limiting the power of tests and making the assessment process more equitable. These strategies include involving diverse groups in the design and administration of tests, considering the consequences of testing, and ensuring that those who develop tests are accountable for their outcomes. Shohamy emphasizes the importance of protecting test-takers' rights, advocating for more participatory and collaborative models of assessment that allow for a broader range of voices and knowledge to be included in the testing process. Through the

application of critical language testing (CLT), Shohamy calls for a shift towards assessments that are democratic, inclusive, and focused on minimizing exclusion and discrimination.

Green (2007) did a review study and in the book *IELTS Washback in Context* examined the fairness of the IELTS Writing task by exploring the test's washback effects, or how the test influences teaching and learning. The book investigates whether IELTS-specific preparation courses provide students with a fair opportunity to improve their writing skills, or if these courses focus too narrowly on test-taking strategies. This raises concerns not only about validity, whether the test truly measures academic writing ability, but also about fairness, as students who cannot access effective preparation may be disadvantaged. Green's research, based on data from 2002 to 2004, assesses whether preparation for IELTS truly reflects the writing demands students will face in higher education, raising questions about the fairness and validity of the test as a measure of readiness for academic environments.

Green's findings reveal significant fairness issues within the IELTS preparation framework. While IELTS-specific courses provide a slight advantage in test scores, they often fail to equip students with the broader writing skills necessary for success in academic contexts. This disconnect suggests that the IELTS test may not fairly represent the full range of writing abilities required in higher education. Furthermore, the research shows that short, intensive IELTS courses may unfairly benefit certain students, particularly those with lower initial scores, while higher-achieving students may not experience the same improvements, indicating unequal benefits from test preparation.

The other source related to fairness in assessments is Fulcher's *Practical Language Testing* (2013) which offers a comprehensive analysis of language testing practices, focusing on the

design, implementation, and implications of standardized assessments such as IELTS and TOEFL. The book addresses key elements like test design cycles, scoring models, test fairness, and the reliability of assessments. While it discusses fairness, especially in relation to rater variability in writing and speaking tasks, it is not solely centered on IELTS.

In a comprehensive empirical research, Hamid et al. (2019) investigated IELTS test-takers' opinions of fairness. The article explores the views of IELTS test-takers regarding the fairness, justice, and validity of the test. Based on survey responses from 430 participants across 49 countries, the study highlights mixed perceptions. While some test-takers considered IELTS necessary and somewhat fair for assessing English proficiency, a significant number questioned whether their scores accurately reflected their language abilities. Concerns were raised about the test's fairness, particularly in relation to its preference for "native" varieties of English, and its role as a gatekeeper for immigration and educational opportunities. Many participants felt that the test did not fully account for different cultural and linguistic backgrounds, thereby disadvantaging non-native speakers.

Moreover, the economic implications of the test were a central concern, with many participants criticizing the high costs of retaking the test and the two-year score validity period. They viewed IELTS as profit-driven, with its policies creating undue financial burdens on test-takers. Inconsistencies in test results across multiple sittings also raised concerns about the reliability and fairness of the test, leading to frustrations over repeated attempts to meet score requirements. The article underscores the need for a more socially responsive approach to standardized language testing, urging greater consideration of test-takers' experiences in test design and administration.

While Hamid et al. (2019) focus on fairness in IELTS, their study is based solely on quantitative survey data. In contrast, our study uses a mixed methods approach, combining quantitative data with qualitative insights from interviews with educators, examiners, and test-takers. This provides a more comprehensive view of fairness, capturing subjective experiences and contextual nuances that Hamid et al.'s study, though valuable, does not fully address.

Another study by Uysal (2010) critically reviewed the IELTS writing test, focusing on key concerns regarding its reliability and validity, particularly because of the test's widespread use in making crucial decisions about test-takers, such as university admissions. Reliability concerns centered on single marking by examiners (the responses are only scored by one rater), which was argued to be insufficient for such a high-stakes exam. Uysal suggested that multiple raters should be involved to improve reliability.

Validity issues were also explored, particularly concerning how well the IELTS writing tasks reflected real-world academic writing tasks, especially in the UK and Australian contexts. Uysal noted that IELTS writing tasks, particularly Task 2, did not fully align with academic and professional genres, and recommended the inclusion of integrated reading-writing tasks to increase authenticity. The article further critiqued the test's claim to assess international English, arguing that its focus was limited to inner-circle varieties of English, neglecting rhetorical conventions from diverse linguistic backgrounds. Uysal concluded by recommending further research into cultural biases, the comparability of test tasks, and the development of a truly international English construct for the writing section.

Another relevant empirical study is the research by Arefsadr and Babaii (2023). In the article *"Let Their Voices be Heard: IELTS Candidates' Problems with the IELTS Academic Writing Test"* they investigated and reported the challenges that IELTS candidates face,

particularly focusing on the IELTS Academic Writing Test. The study involved interviewing 10 Iranian IELTS candidates to understand why writing consistently appears as the lowest-scored skill compared to other sections of the exam. Four primary issues were identified through thematic analysis of the interview data: insufficient time, unclear and difficult-to-understand task instructions, "distant" topics, and the overvaluation of advanced vocabulary and grammar in the scoring system. These findings suggest that candidates' lower writing scores may not always reflect a lack of proficiency but could stem from the design of the writing test itself, particularly in the way it values certain linguistic features.

The study emphasized the need to rethink some of the components of the IELTS Writing Test to make it more equitable and accessible to a diverse group of test-takers. For example, candidates expressed frustration with distant, too broad or academic topics, insufficient time to write thoughtful responses, and the perceived pressure to use advanced vocabulary and grammar that may not always be necessary for clear academic writing. These findings raise important questions about test fairness and suggest potential areas for improvement, particularly in ensuring that the test measures writing proficiency without introducing unnecessary hurdles that could disadvantage certain groups of test-takers.

Role of Technology in Language Assessment

In recent decades, technology has significantly influenced practices in language assessment in terms of test construction, delivery, and scoring. This has been majorly propelled by the growing need for testing solutions that are both efficient and scalable, while at the same time being user-friendly for the increasing number of test-takers worldwide. While paper-and-pencil-based traditional tests were the mainstay in assessing languages, there is a fast or gradual replacement being facilitated with Computer-Based Testing (CBT) systems and automatic scoring. These

technological advancements include such pluses as the much faster processing of results, increased consistency in scoring, and the inclusion of multimedia elements that lead to a more holistic evaluation of language skills (Kim & Lopez, 2022).

Digital technology changes the requirement for a shift from traditional to digital assessments, especially relevant for the fairness and accessibility of high-stakes language tests like IELTS, guiding the research questions of this thesis. Day by day, such testing configurations are being assimilated (Erickson & Tholin, 2022). It is consequently urgent to explore how these shifts will affect assessment equity among test-takers from such vastly different linguistic, cultural, and socio-economic backgrounds. This discussion will now move towards considering the impacts of technological advancements in language testing in relation to increasing, or alternatively, possibly further entrenching, issues of fairness and accessibility.

The discussion has drawn on a wide range of existing literature to offer the fullest current picture of the status of integration of technology into assessment practices as a means of either levelling the playing field for test-takers or introducing new bias and inequality. The subsequent sections will address specific issues of technology's function in language testing, such as the benefit and challenges posed by computer-based testing, the impact of automated scoring systems, and the relentless requirement for technological advancement and research in an area that is still far from being stabilized (Gokturk & Tsagari, 2022).

Computer-Based Testing (CBT)

Computer-Based Testing defines practice of delivering tests electronically and not in the conventional way of using paper and pen methods. This move towards progress for modern language assessment avails a test setting that is far much flexible and dynamic. In the case of CBT,

students will sit for assessments using computers, most times in test centers, or at home, and the responses are directly recorded and submitted electronically. Some of the most famous tests of English language proficiency include the TOEFL iBT, IELTS, and Duolingo English Tests, all of which have adopted or incorporated CBT to some extent. The introduction of CBT introduces a critical linguistic assessment methodology change and brings both opportunities and challenges in the assessment arena into practice.

The first key benefit is that CBT allows flexibility in administering the tests. It enables the test taker to do it more often and at different places, giving those wanting to take these tests a wider margin of where and when to test. This makes it so flexible and is most rewarding in meeting the demands of a global test population, even among less accessible countries or remote areas with such services thinly laid out, like most of the developing countries today (Chapelle & Douglas, 2006). Additionally, CBT aids in logistics, which reduce the magnitude of the materials needed for testing, besides simplifying the overall process of administering tests and sending results. Moreover, CBT allows for construct-irrelevant variables such as messy and illegible handwriting to be minimized.

An important advantage of CBT allows incorporating multi-media segments in a test is that, unlike the traditional administration of the paper and pencil test, these newer versions incorporate a way to embed easily audio, video, and interactive tasks to offer a concise summary assessment of language skills on the part of a test-taker. That multi-modal medium allows a valid assessment of test-takers' listening, speaking, and writing skills since its similarity to real-life communication experiences is very high (Wang et al., 2007).

Admittedly, CBT has both its advantages and its challenges, particularly on the line of fairness and accessibility of the test. The overwhelming challenge however is in what is termed

digital divide, which is a gap between people who do and do not have access to technology. In that way test-takers who come from lower socio-economic backgrounds or from regions relatively poor in access to technology will be disadvantaged in this case. These test takers might not have prerequisite level of digital literacy, or they might not even have access to reliable internet access, and either one of these conditions may negatively affect their score in a CBT.

Moreover, digital literacy among candidates may also be different, which may bring problems for access or completion within computer-based assessment tools. Some test-takers do have good skills in keyboarding, using computers and digital interfaces, but others do not, especially those test-takers who have been used to traditional paper-and-pencil tests. It can bring about differences that are not really related to the candidate's actual language ability, but to a candidate's familiarity with technology itself (Wang et al., 2007).

It also results in issues of access on racial and socioeconomic lines. As an illustration, candidates who are from rural or remote regions face challenges accessing CBT centers. Or, where this is conducted remotely, the candidates could face challenges of internet reliability. In both cases, it opens gates to unequal access to language test resources, which brings forward the fairness of the CBT system in testing candidates on a level playing field.

Automated Scoring Systems

Automated scoring systems use artificial intelligence and natural language processing technologies to score test-takers' written and spoken responses to test questions. They are predominantly developed to process vast quantities of linguistic data and provide prompt, consistent and objective scoring. Notable types of automated scoring systems are the ETS e-rater for the TOEFL iBT and the Duolingo English Test.

Automatic scoring systems offer the advantages of consistent and objective assessments. Unlike human raters who may introduce bias through subjectivity or scorer fatigue, automatic systems apply the same criteria to all responses, ensuring uniform ratings for each test taker. This reduces the kind of score variation that can happen with human raters, and there will be reliability in differentiating a measure of language ability (Shermis & Hamner, 2013).

Moreover, these automated scoring mechanisms would effectively eliminate the forms of biases through human judgment. Since algorithms do most of the rating, the effects of a rater's personal likes and dislikes or any other kind of prejudicial thinking are reduced. Such objectivity becomes crucial when it comes to language testing that carries high stakes, as even minor biases can have serious consequences for test-takers (Shermis & Hamner, 2013).

However, automated scoring systems have their own concerns. One central problem is the potential biases that algorithms themselves could input. The training data for these algorithms will influence performance, providing room for bias in scoring. For instance, the authors point out that if the training data predominantly comprises answers from a particular demographic group, the algorithm might favor a certain type of language or speaking style for that particular group at expense of test-takers who belong to a different demographic. A much more critical issue may be how effectively automated systems would capture all the complex linguistic features, such as creativity, humor, or cultural topics. While automated systems are great estimators of more mechanical features, such as grammar and syntax, the sensitive and complex features of communication can prove to be a daunting task to judge accurately by these systems. This limitation casts doubt on the validity of automated scores with respect to types of tasks that call for deeper contextual and cultural understanding (Shermis & Burstein, 2013).

It is important to note that while automated feedback systems are commonly used in preparation for the IELTS exam, particularly to provide students with immediate feedback on their writing, the actual IELTS writing test, including Task 2, is still scored by human raters. Even in the computerized version of the test, human raters evaluate written responses. However, technology aids in providing a clearer representation of typed answers, which helps reduce issues related to legibility that may occur with handwritten responses. However, the use of computers in the writing task may present challenges, such as unfamiliarity with typing for some test-takers, leading to slower responses or increased anxiety (Weir *et al.,* 2007). There are also concerns that the use of computerized tests may create inequalities for those less comfortable with technology or typing, as frequency of word processing has been shown to affect performance. These concerns point to the need for greater transparency and training in human scoring processes to ensure fairness. Additionally, while technology helps in delivering tests efficiently, human involvement in the evaluation remains crucial in addressing deeper contextual or cultural nuances in writing (Weir *et al*., 2007).

Impact of Technology on Fairness

Digital Divide and Equity. The above discussion demonstrates that the introduction of computer-based testing has made language assessment easy, effective, and productive, though along with these there are major issues of access to technology. This points toward the digital divide, pertaining to the gap that exists between the people who have access to modern information and communication technologies and those who do not. The digital divide has been well documented to place test-takers, whose background is underprivileged or whose region possesses less

infrastructural development in terms of technology, at a disadvantage in CBT settings. These individuals may lack or have limited access to a computer, high internet connection, or up-to-date software, which would disfavor them to show high levels of performance on computer-based testing individual computers.

The following intervention measures and suggestions have been previously made to respond to the above disparities. One potential way of tackling this is by ensuring testing centers are adequately resourced, and then providing sufficient access to such resources for test-takers from disadvantaged social backgrounds in advance of the examination. A body like IELTS could provide pre-exam free or low-cost to computer labs or, on the other hand, allow evaluation methods for those also not having access to technology. By these measures, closing the digital divide would ensure an equal opportunity to succeed with digital tools and environments for all test-takers, across socio-political and geographical lines.

Digital Literacy. More than access to the technology, digital literacy—that is, being able to use digital tools and platforms effectively—is also an important consideration in the fairness of the technology-enhanced language assessment. Digital literacy of the test-takers does make a difference to test outcomes. Candidates who are more adapted to digital interfaces and typing may find computer-based testing much easier and less time-consuming, while those not so familiar may struggle with the mechanics of the test rather than the content itself. This inconsistency can grotesquely tilt the results of the test and establish a partiality in favour of the more technologically inclined section, while the language skills might not necessarily be better for them (Lindner & Greiff, 2023).

In these lines, it should be depicted that digital literacy should be induced as part of the testing preparedness. These testing organizations could offer tutorials or practice sessions in which

the candidates are made familiar with the digital format of the test so that all candidates acquire the necessary skills to navigate in the computer-based environment. This training can come online or through workshops in testing centers, which would create fairness and make the test outcomes a reflective measure of language ability, not digital ability (Wang et al., 2008).

Cultural Bias in Tech-enabled Testing

Algorithmic Bias. Automated scoring systems, often used in IELTS preparation, are designed to provide objective feedback on students' writing. However, concerns about algorithmic bias arise when these systems are trained predominantly on essays from native English speakers. This could result in these systems favoring specific linguistic structures or writing styles that are common in native speakers, potentially disadvantaging non-native speakers whose writing may deviate from these 'standard' forms, despite being equally effective in communication (Shermis & Burstein, 2013).

For IELTS, this raises the question of fairness, as test-takers from diverse cultural backgrounds might be penalized by such biases in preparation tools. While the actual scoring of IELTS writing tasks is still done by human raters, the integration of AI-assisted feedback in preparation highlights the importance of developing scoring systems trained on a large and diverse dataset, reflective of the global variability of English. Additionally, continuous monitoring and updating of these systems are crucial to minimizing emerging biases and ensuring fairer outcomes in technology-enhanced language testing

Design in Test Prompts. Cultural biases can also be found in the design of test prompts in language assessments, including computer-based tests. While Shohamy (2001a) does not specifically focus

on computerized testing, her critique of cultural bias in assessment can be applied broadly to test design and fairness issues across different formats. Test content should be designed that is neutral to the culture of any individual or causing any cultural problems in the understanding and interpretation of test content. Test prompts that are grounded in a specific cultural context that is foreign to some test takers lead to an inability to relate to the presented material, thus resulting in low performance. For example, the so called cultural-specificity prompt might render test-takers from other cultures unable to familiarize themselves with the matter, consequently scoring low points that do not represent their actual language competencies.

Tasks toward cultural sensitivity in test design should include a thorough review of the test content that can identify and eliminate the prompts containing cultural bias. Experts in linguistics and culture should be part of the task force since they will offer clear insights on some of the potential bias that the test items may elicit. Furthermore, piloting of prompts with representative groups of examinees can be used to eliminate any cultural biases that may be present and sorted out prior to the administering of the test to a larger group of examinees.

Theoretical Framework for the Study

The present study is grounded in established theories and frameworks on test fairness, bias, and validity in language assessments. Central to this analysis is the Test Fairness Framework (TFF) proposed by Kunnan (2004), which provides a comprehensive lens for evaluating the fairness of high-stakes language tests like IELTS. Kunnan's framework emphasizes five key dimensions of fairness: access, administration, absence of bias, social consequences, and validity. These elements guide the investigation into whether the IELTS writing test equitably assesses all test-takers,

irrespective of their linguistic or cultural backgrounds. For example, for the developed questionnaire, five dimensions of Kunnan's Test Fairness Framework are addressed. Under "Access," questions focus on issues such as the financial costs of the tests, which may limit equal opportunities for candidates. For "Administration," factors like timing is considered, examining whether the allocated time for the writing tasks is sufficient for all test-takers to perform optimally. The "Absence of Bias" is explored by asking participants whether they believe the writing prompts and scoring criteria are culturally neutral and free from any unfair advantages or disadvantages. The "Social Consequences" dimension is captured by questions investigating the test's impact on life opportunities, such as immigration, education, or job prospects, highlighting how IELTS results influence the social and professional lives of candidates. Lastly, "Validity" is covered through questions evaluating whether the test measures the true language abilities of test-takers, ensuring that it accurately reflects their proficiency in real-world language use.

Additionally, Bachman and Palmer's (1996) notions of construct validity and authenticity in language assessment are essential. Their work emphasizes the importance of ensuring that test tasks reflect real-world language use and that scoring criteria align with the intended constructs. This study uses Bachman's principles to examine how well the IELTS writing tasks align with academic writing expectations and whether scoring practices remain unbiased across diverse groups.

Fairness and cultural bias in language testing have also been explored by Shohamy (2001a, 2001b), who highlights the role of power and politics in the design and implementation of language assessments. Shohamy's critique emphasizes the need for test designers to actively address potential sources of cultural bias in test prompts and criteria. This study draws on Shohamy's

arguments to examine whether the IELTS writing assessment may inadvertently privilege certain cultural or linguistic backgrounds.

Lastly, the increasing role of technology in language assessment is examined through the lens of Hamp-Lyons' (2016) work on automated scoring and computer-based assessments. While technological innovations offer new avenues for standardization, they also introduce concerns about accessibility and fairness, particularly for test-takers from less technologically proficient backgrounds. The current study integrates these concerns into its investigation of how technology impacts the fairness of the IELTS writing test, particularly for candidates completing the computer-based version.

Together, these theoretical perspectives provide a robust framework for analyzing the fairness of the IELTS writing test from multiple angles.

Conclusion

This review of literature provides an extensive overview of the important concerns related to the fairness of IELTS writing test, Task 2. In addition to technical validity and wider societal implications, the review emphasizes ethical necessity and varied nature of fairness in language testing. Despite the IELTS's widespread recognition and influence, issues with cultural bias, examiner subjectivity and uneven access to preparation materials continue to be raised by the literature. These elements could affect test's validity and fairness especially for applicants with different language and cultural backgrounds.

While there are potential advantages in consistency and efficiency from use of technology in language evaluation, there are also new issues with digital literacy and accessibility. To make

sure that new developments in technology do not make already existing disparities worse, continued attention must be paid to digital divide and any biases in automated scoring systems. Furthermore, to ensure fairness and justice for every test-taker, the theoretical frameworks and empirical investigations discussed in this chapter highlight necessity of ongoing assessment and improvement of language testing procedures.

Chapter 3
Methodology

Research Design

This study used both qualitative and quantitative methodologies to investigate fairness of IELTS writing assessments. This approach was ideal for addressing issues of fairness in high stakes language tests because, by integrating qualitative insights from interviews with quantitative data from questionnaires, this study sought to provide a comprehensive investigation of fairness as viewed by different stakeholders participating in the IELTS writing test. Simultaneous emphasis on subjective experiences and empirical information increased validity of findings and allowed for a more nuanced analysis of the study issues.

The qualitative component of the study included semi-structured interviews that were intended to provide an in-depth description of participants' lived experiences and perceptions of fairness. This approach gave flexibility and customization based on interviewees' responses and allowed the researcher to capture each participant's unique and contextualized experiences. Interviews were designed to acquire detailed information about participants' perspectives on fairness in IELTS writing assessment, the obstacles they encountered, and the impact of technology on assessment fairness. The quantitative component used a questionnaire to gather information from larger group of participants. This enables the researcher to measure opinions of fairness while also identifying broader patterns and trends. The questionnaire was developed to collect data on critical problems related to fairness in IELTS writing test. The use of both qualitative and quantitative methodologies enabled data triangulation which improved the validity and trustworthiness of the research results. This mixed methods approach was consistent with

research questions, which aimed to investigate various aspects of fairness in IELTS writing assessment.

The combination of qualitative and quantitative data gave a thorough grasp of issues at hand and allowed the researcher to address both individual participants' subjective experiences and the broader trends revealed by the data.

Participants and Sampling

Sampling approach

This study used purposive sampling. It is also known as judgmental or selective sampling. This kind of sampling is a type of non-probability sampling and in which participants are chosen based on their relevance to the research goals. This sampling method is especially helpful when researcher has to collect lots of data from participants who have similar characteristics or experiences. Purposive sampling for example, is often used in research that focuses on rare populations such as those with specific skills or individuals who have experienced unusual events. This study used purposive sampling to select people who had direct engagement in IELTS writing assessment either as test takers or educators and examiners. The researcher selected participants based on their experiences with IELTS writing test and ensured sample was
relevant to study's aims at fairness in tests (Palinkas et al., 2015).

Participant Selection

The study involved 30 participants, divided into two groups: 15 IELTS test takers and 15 IELTS educators or examiners (10 educators and 5 examiners). Since many of the examiners were also IELTS educators, we combined these roles into one group. This grouping ensured that the

perspectives of both test takers and evaluators were well-represented, providing a comprehensive understanding of fairness in IELTS writing assessments. Participants were chosen based on the following inclusion criteria. They were at least 18 years old. They had direct experience with IELTS writing test as test takers, educators or examiners. They were able and they were willing to offer informed consent to participate in study (See Appendix C). Participants also had the necessary English language skills to participate in interviews and complete questionnaires, which were administered in English. Individuals under the age of 18, those with no direct experience with the IELTS writing assessment, and those unable to offer informed consent or participate in the study owing to language hurdles or cognitive limitations were also excluded. No participants were specifically chosen based on characteristics such as ethnicity, gender or socioeconomic status. In other words, focus was on their roles within IELTS assessment process.

Recruitment of participants was done through online and offline strategies. In this process, information about the study was disseminated through academic and professional networks, social media platforms and online forums for English language education and IELTS testing. This recruitment technique guaranteed that study included a variety of participants from different professional and cultural backgrounds, which was critical for understanding the complexity of fairness in a global assessment situation such as the IELTS (See Appendix D).

Data Collection Methods

In this study, the required data were gathered using two main methods: questionnaires and semi-structured interviews. Using these in combination led to collecting both quantitative and qualitative data which together gave a comprehensive picture about fairness in IELTS writing tests.

Questionnaires. All 30 participants completed questionnaires. The questionnaire was developed to collect quantitative data on perceptions of participants on fairness in IELTS writing assessments. The questionnaire can be found in Appendix B.

The design and development of the questionnaire for this study was an essential step toward ensuring that the research objectives were fully addressed. This process was guided by the research problems and theoretical frameworks that underpin this study, which include Kunnan's Test Fairness Framework (2004) and fairness considerations in high-stakes language examinations (Kunnan, 2010). While developing the questionnaire, careful attention was given to ensure that questionnaire met validity criteria in that it accurately examined participants' perceptions of fairness in IELTS writing test.

The primary focus of the questionnaire was to explore how educators and examiners perceive fairness of the IELTS writing test, the challenges test-takers face and how technology influences fairness in scoring. The design of the questionnaire was thus aligned with these core research questions to ensure that the data collected would directly address these issues. The questionnaire's development was enhanced by basing it on theoretical constructs from Kunnan's framework (2004). This meticulous alignment provided construct validity by making sure questions reflected theoretical aspects of fairness as intended and the questionnaire measured the essential concepts relevant to the study (Brown, 2000).

A small number of individuals took part in a pilot test before the questionnaire was finalized. This was a step in refining the questions, assuring clarity, and enhancing the general flow of the questionnaire. The pilot test feedback enabled linguistic adjustments in specific questions as well as the elimination of ambiguities, all of which contributed to face validity. It also contributed to content validity by confirming that all relevant themes were included in final version

of questionnaire as well as reliability by ensuring that participants understood the questions clearly and consistently.

The questionnaire consisted of closed-ended questions, mostly utilizing Likert scales to test participants' level of agreement with statements about fairness in IELTS writing tests. Likert scales are widely used in social science research for measuring attitudes and opinions, which helped improve construct validity by allowing subjective perceptions to be quantified reliably (Boone & Boone, 2012).

To improve reliability, questionnaire included multiple items that addressed each significant issue such as scoring fairness. It provided internal consistency by cross-checking responses to related questions (Field, 2013). The Likert scale was used in this study because it is a simple, effective way to measure attitudes and perceptions. It allows respondents easily express how strongly they agree or disagree with a statement and makes it ideal for collecting structured data on fairness (Likert, 1932; Jamieson, 2004). Likert scales produce reliable and consistent results by measuring the same concept across multiple related items which allow for a consistency in response gathering and analysis. This reliability aids in identification of themes in the data and ensures accuracy and makes it excellent for studying complex issues such as fairness (Boone & Boone, 2012; Field, 2013).

Finally, during the questionnaire preparation process, ethical considerations were carefully addressed. The responses which the participants gave were kept confidential and their participation was voluntary with informed consent provided before study. Ethical practices not only ensured that the study fulfilled ethical requirements but they also improved overall validity of data by creating a context in which participants felt comfortable offering honest and accurate responses (Bryman & Bell, 2011).

Questionnaires were distributed via email and then participants were given two weeks to complete and return questionnaires, with follow-up reminders as needed.

Interviews. In current study, we conducted semi-structured interviews with some of participants from both test-takers and educators/examiners. In total, ten participants (five persons for each role) were chosen for in depth interviews which allowed for a deeper look at their personal experiences and perceptions of fairness in IELTS writing tests. In addition, semi-structured nature of interviews caused flexibility in the questioning process. It enabled the researcher to probe deeper in issues raised by participants. It also ensured that key topics related to research questions were covered.

The interview guide (Appendix B) was carefully designed to align with the study's research questions. The guide had open-ended questions that invited participants to discuss the overall fairness of IELTS writing testing, the difficulties they encountered and their perspectives on role of technology in ensuring or undermining fairness. For example, participants were invited to consider whether they thought IELTS writing test was fair to all test takers regardless of background and to address any specific challenges they or others had encountered during assessment process. Interviews were conducted via Zoom to accommodate geographical locations and preferences of participants. Each interview lasted approximately 30 minutes and was recorded with participants' consent. The interviews were transcribed and transcripts were used as primary data source for qualitative analysis.

Data Analysis

The data analysis technique was divided into two parts: qualitative analysis of interview transcripts and quantitative analysis of questionnaire responses.

We utilized thematic analysis to investigate qualitative data collected through semi-structured interviews. We followed Braun and Clarke's (2006) six-stage process. Thematic analysis is a versatile and widely used technique for identifying, understanding, and presenting patterns or themes in qualitative data. It enabled the researcher to go beyond simply documenting the data and instead analyze and interpret the underlying meanings and implications. First phase of analysis entailed getting familiar with data by reading and rereading interview transcripts. This stage was important for acquiring a thorough knowledge of data and discovering early patterns and insights. Second phase entailed creating initial codes by carefully categorizing data into distinct groupings. Coding was done manually with codes assigned to specific sections of the text that were relevant to the research topics. During the third phase, the basic codes were organized into probable themes. Themes indicated wider patterns of relevance throughout the data and were derived from the clustering of similar codes. In the fourth phase, themes were reviewed and modified to ensure that they accurately represented data and addressed study questions. Any themes that were not compatible with entire dataset were redefined or removed. During fifth phase, themes were defined and labeled and each theme was analyzed. The final step involved writing analysis, combining themes into a cohesive narrative that addressed research goals related to fairness of IELTS writing. Quotes obtained from interviews were used to explain themes and back up findings.

In the quantitative analysis of study, by means of a semi-structured questionnaire, we collected data from 30 participants which were 15 educators/examiners and 15 test-takers. To ensure reliability and validity of instrument, Cronbach’s alpha was employed which revealed moderate to high internal consistency across various sections of questionnaire (0.613, 0.644,

0.871). Moreover, we had an expert validation. It confirmed content validity of questionnaire while aligning it with established theories of fairness in language testing.

For statistical analysis we employed both descriptive and inferential methods. Descriptive statistics such as means, standard deviations and frequency distributions provided a summary of participant responses. Inferential statistics utilized Wilcoxon Signed-Rank Test and t-tests. Results of quantitative analysis were presented in tables and figures and they were accompanied by a narrative interpretation of findings. The quantitative data were used to support and complement qualitative findings and provided a more comprehensive understanding of fairness in the IELTS writing assessment.

Following a preliminary analysis of the interview data, we used member checking (McKim, 2023) to enhance the trustworthiness of the study. This involved sending each participant a summary of the most salient findings and emergent themes. Comments were invited on whether these interpretations reflected their experiences and perceptions. This helped assurances that emergent findings were participant-based and not solely driven by the researcher.

Ethical Considerations

Ethical guidelines which were followed during the study were according to those stipulated by Concordia University. Research ethical approval was given through the Ethics Committee at Concordia University and all activities were conducted in accordance with approved protocol.

The informed consent of the participants was obtained before they participated in this study. An information sheet explaining the purpose of the study, procedures to be followed, possible risks and benefits and their rights as participants such as the right to withdraw at any stage without penalty was given to participants. Participants taking part either in interviews or in the completion

of questionnaires needed consent forms. Consent sheets were stored in a locked cabinet for security; all data were anonymized to maintain privacy and confidentiality among participants. All data including interview transcripts and questionnaire responses were stored on a password-protected computer. Any physical documents were kept in a locked cabinet. Data will be retained for five years after completion of study; after this time, it will be securely destroyed. No identifying information was included in any publication or presentation which resulted from this research.

Limitations

The findings provided a broader understanding of IELTS writing assessments with regard to fairness; however, there were various limitations. To begin with, although a sample size of 30 participants is adequate for the research design, this may limit the generalization of findings to the wider population of both IELTS candidates and educators. Reliance on purposive sampling in this study also had the implication that findings could not be generalized to represent all IELTS writing assessment stakeholders. Because of this, the data extracted from semi-structured interviews were bound to be subjective and may have reflected personal biases and perspectives held by participants and the researcher. While thematic analysis provided a structured approach toward the analysis of qualitative data, the interpretation by the researcher may also become susceptible to biases. Another limitation it presents is that thematic analysis does not provide an in-depth analysis compared to other approaches such as content analysis (Braun & Clarke, 2006).

Chapter 4

Results and Discussion

This chapter reports the findings from quantitative and qualitative analyses in response to the research questions in this study. The data were collected from two groups: educators/examiners and test-takers, and it was evaluated using both descriptive and inferential statistics.

Quantitative Analysis

The quantitative analysis is based on data from a questionnaire distributed to 30 people including 15 educators/examiners and 15 test takers. Participants reported a variety of first languages, such as Persian, Chinese, English, Arabic, Turkish, Pashto, Vietnamese, Indonesian, Thai, Russian, and Kurdish. This section includes the reliability analysis, descriptive statistics and inferential statistics.

Questionnaire Validation and Reliability

The reliability and validity of the questionnaire used in this study were evaluated using two complimentary methods: statistical reliability analysis with Cronbach's alpha and expert validation. These methodologies were used to verify that the questionnaire accurately examined the main dimensions of the study.

Cronbach's alpha was computed for every part of the questionnaire to assess the items' internal consistency (see Table 1).

Table 1

Cronbach's Alpha for Questionnaire Section

Section	Number of Items	Cronbach's Alpha
Section A (Educators/Examiners)	6	0.613
Section B (Test-Takers)	6	0.644
Section C (Technology)	4	0.871

The results showed that Section A (Educators/Examiners) had a Cronbach's alpha of 0.613, indicating moderate internal consistency. Although this value is slightly below commonly accepted threshold of 0.7, it is considered acceptable in exploratory research and for subjective constructs like perceptions of fairness (Tavakol & Dennick, 2011). Section B (Test-Takers) had a Cronbach's alpha of 0.644 which suggests acceptable internal consistency for items related to test-takers' experiences with IELTS writing assessment. Although slightly below the optimal level, this value is typical for studies with smaller sample sizes and exploratory research in language testing (Field, 2013). Section C, which examined the role of technology in fairness, had a Cronbach's alpha of 0.871, indicating high internal consistency. This strong value demonstrates that the items in this section are highly correlated and consistently measure the same underlying concept—namely, how technology impacts fairness in language testing.

While Cronbach's alpha is effective for assessing internal consistency, it is best combined with qualitative measures like expert validation. In this study, a panel of specialists in language testing reviewed the questionnaire to ensure its content validity particularly given small sample size (Haynes, Richard, & Kubany, 1995). The experts confirmed that items were aligned with established theories of fairness in language testing. They found that Section A effectively

addressed fairness in IELTS writing assessments particularly regarding cultural and linguistic biases. They also validated Section B and confirmed that it addressed key challenges like task difficulty and time constraints consistent with themes in high-stakes language testing. Section C was similarly validated and reflected how technology affects fairness, a topic increasingly discussed in assessment research.

Overall, the combination of Cronbach's alpha and expert validation confirms the questionnaire's reliability and validity. While Cronbach's alpha provided quantitative internal consistency, expert validation ensured content validity, aligning the items with theoretical frameworks and current issues in language testing. Future research with a larger sample size would be beneficial in order to provide more robust validation via exploratory factor analysis; however, we believe that the validation was adequate for such exploratory research.

Descriptive Statistics

Descriptive statistics are used to summarize key characteristics of data including measures such as mean, standard deviation, minimum, maximum and frequency distributions. The following figures present a summary of the responses from educators/examiners and test-takers on various aspects of IELTS writing assessment.

Section A: Perceptions of Fairness in IELTS Writing Assessment (Educators and Examiners) (Questions 5 to 10).

Q5. How clear do you find IELTS writing assessment criteria?

Figure 1

Perceptions of Clarity of Scoring Criteria

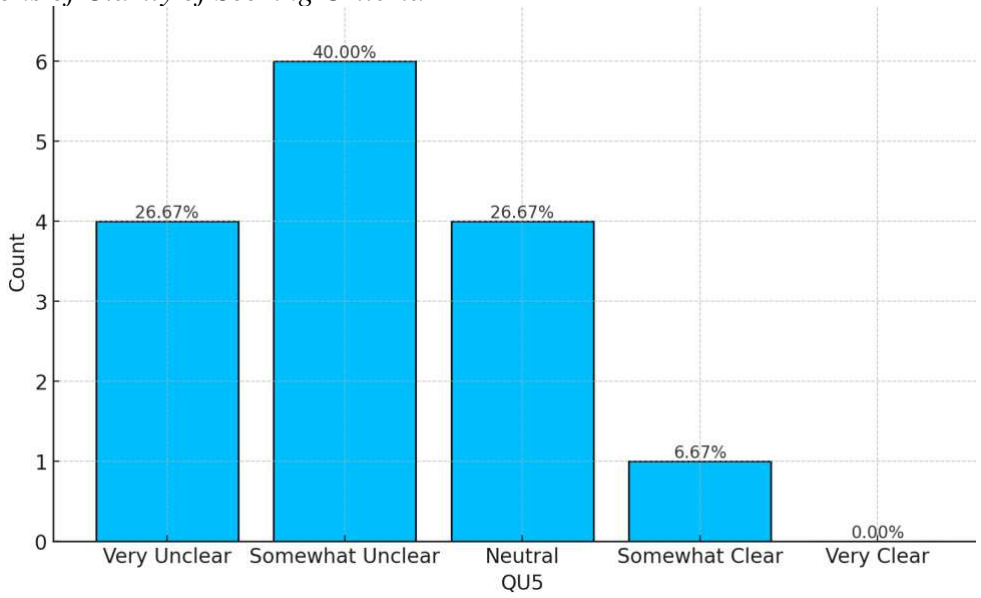
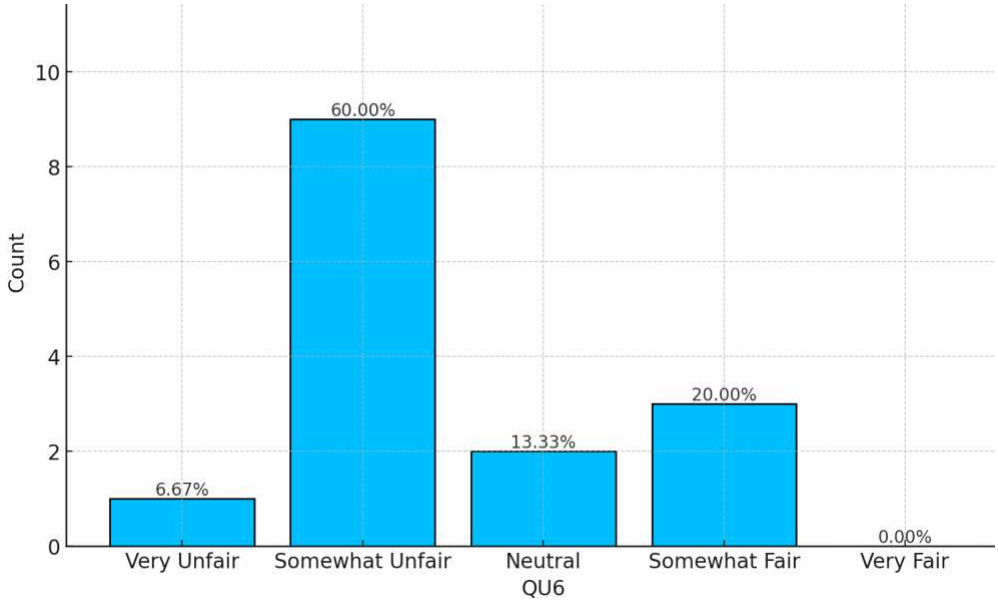


Figure 1 shows that 40% of educators and examiners find IELTS writing assessment criteria somewhat unclear while only 6.67% consider the criteria to be somewhat clear.

Q6. How fair do you think IELTS writing assessment is for test-takers from diverse cultural and linguistic backgrounds?

Figure 2

Perceptions of Fairness for Culturally and Linguistically Diverse Test-Takers

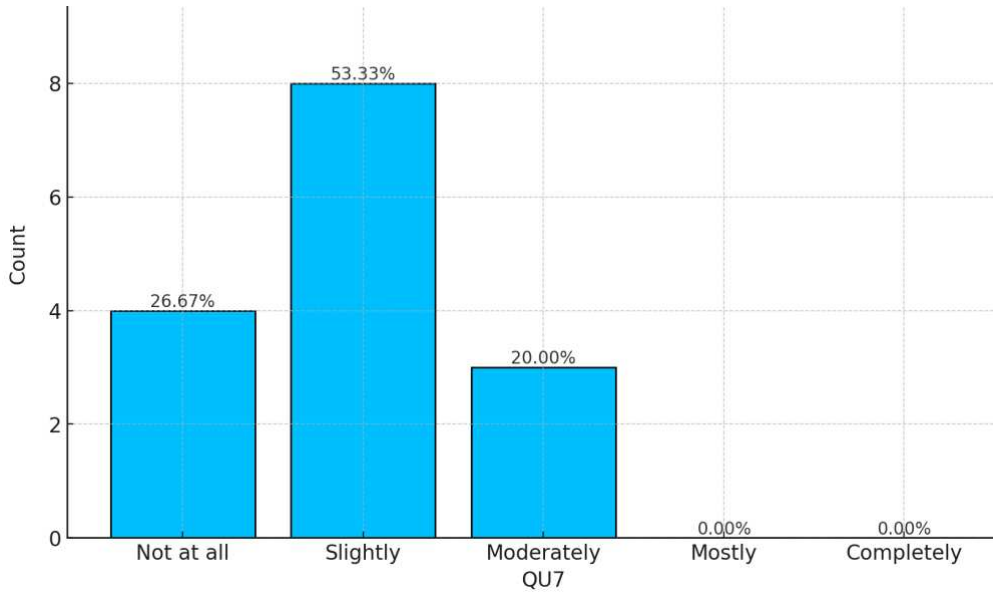


As illustrated in Figure 2, 60% of educators and examiners believe that IELTS writing assessment is somewhat unfair for test-takers from diverse cultural and linguistic backgrounds while only 20% find it somewhat fair.

Q7. To what extent do you believe IELTS writing assessment addresses cultural differences?

Figure 3

Perceptions of How IELTS Writing Assessment Addresses Cultural Differences



According to Figure 3, 53% of respondents believe that IELTS writing assessment only slightly addresses cultural differences while 24.67% feel it does not address them at all. A smaller proportion (21.33%) thinks assessment moderately considers cultural differences.

Q8. How often do you think bias occurs in IELTS writing assessment due to cultural or linguistic differences?

Figure 4

Frequency of Perceived Bias in IELTS Writing Assessment Due to Cultural or Linguistic Differences

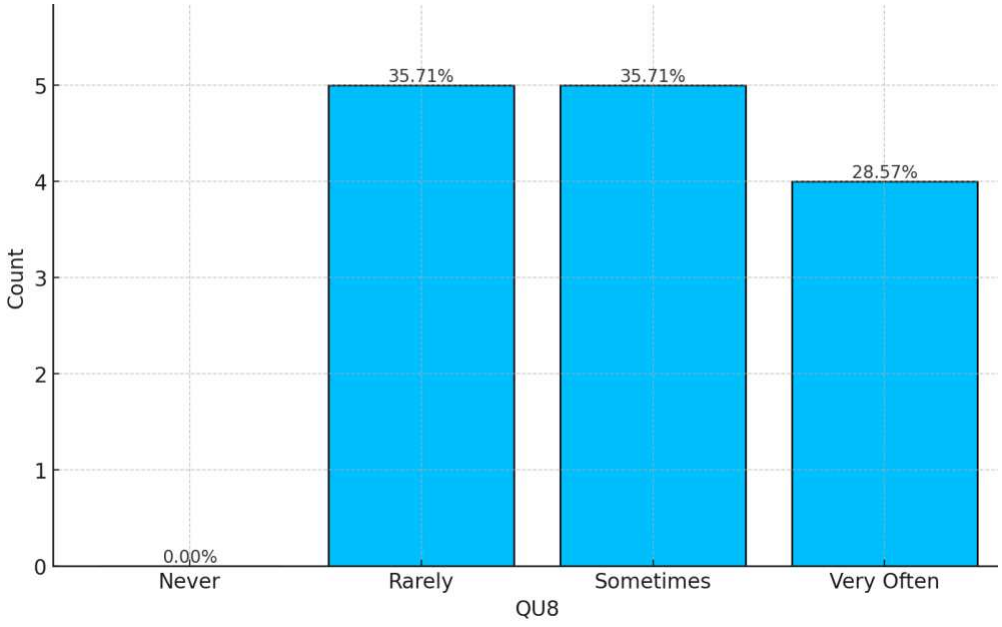
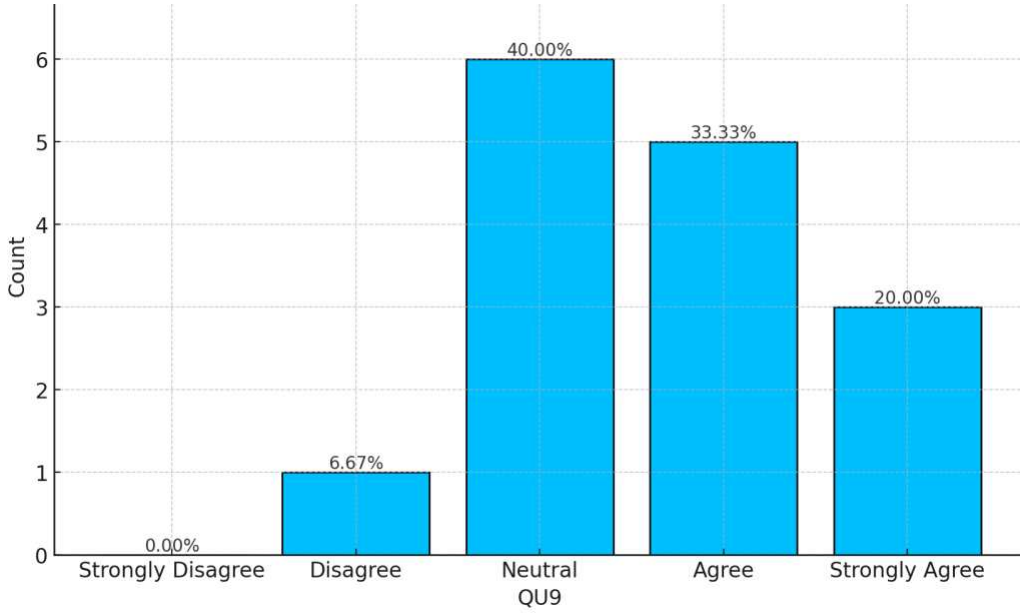


Figure 4 shows that 28.57% of respondents believe that bias due to cultural or linguistic differences occurs very often in IELTS writing assessment while 35.71% think it happens sometimes. An equal percentage (35.71%) believes it occurs rarely with no respondents stating that it never occurs.

Q9. In your opinion, do IELTS writing prompts reflect a Western-centric perspective?

Figure 5

Perceptions of Western-Centric Bias in IELTS Writing Prompts



As shown in Figure 5, 33.33% of respondents agree that IELTS writing prompts reflect a Western-centric perspective and 20% strongly agree. A significant portion of respondents holds neutral opinions while only 6.67% disagree with the statement.

Q 10. Do you believe that emphasis on IELTS writing scores in university admissions, job placements or immigration decisions is fair?

Figure 6

Perceptions of Fairness in Emphasizing IELTS Writing Scores for University Admissions, Job Placements and Immigration Decisions

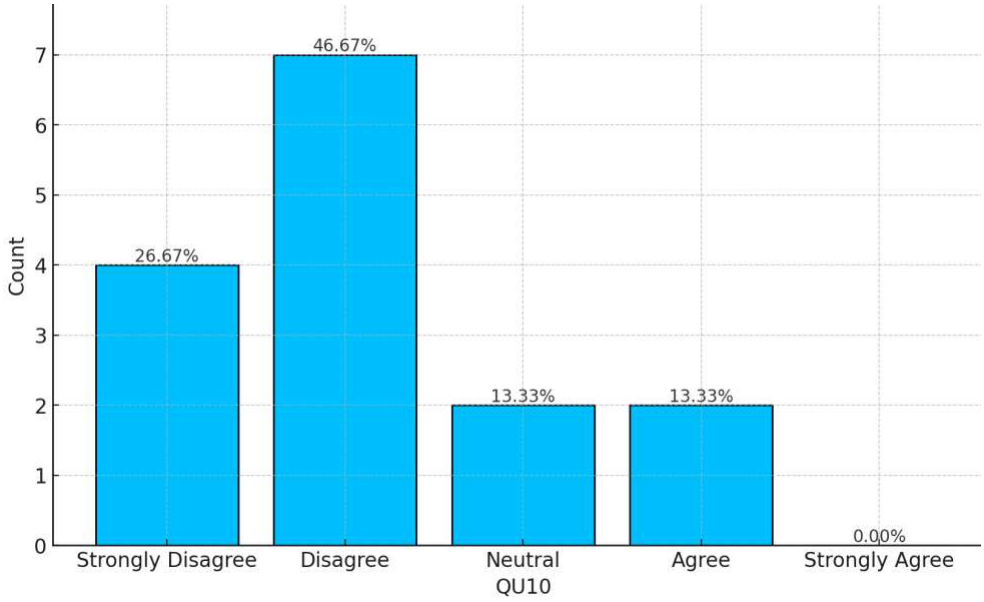


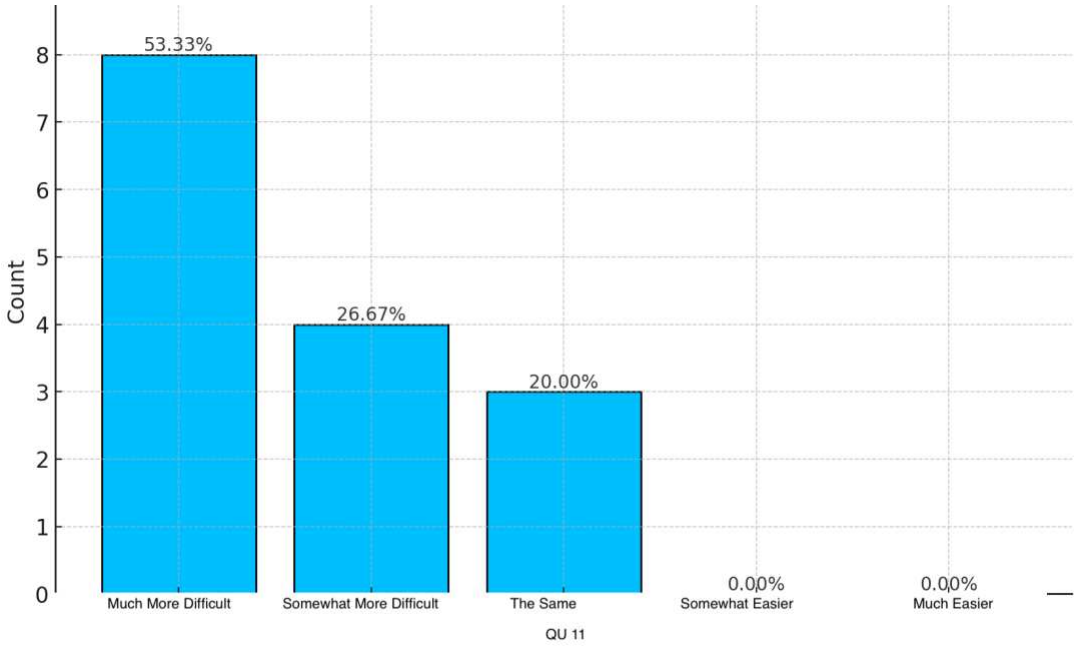
Figure 6 highlights that 46.67% of respondents disagree that emphasis on IELTS writing scores for university admissions, job placements or immigration decisions is fair while 26.67% strongly disagree. Only 13.33% of respondents hold neutral or agreeing views on fairness of IELTS writing scores in these decisions.

Section B: Challenges Faced by Test Takers (Questions 11 to 16)

Q 11. How difficult do you find the IELTS writing assessment tasks compared to real-world writing requirements?

Figure 7

Perceptions of Difficulty of IELTS Writing Test Compared to Real-World Writing Requirements

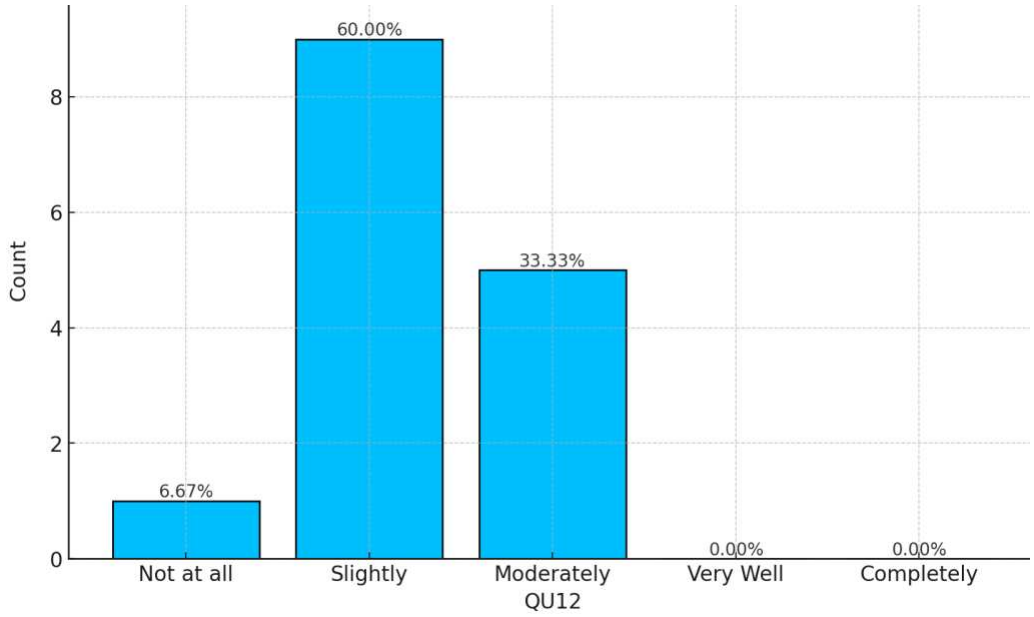


As depicted in Figure 7, 53.33% of test-takers find the IELTS writing assessment tasks to be much more difficult than real-world writing requirements while 20% consider the tasks to be about the same. 26.67% of respondents believe tasks are somewhat more difficult than real-world writing.

Q 12. How clearly do you understand scoring criteria used in the IELTS writing assessment?

Figure 8

Perceptions of Scoring Criteria in IELTS Writing Assessment

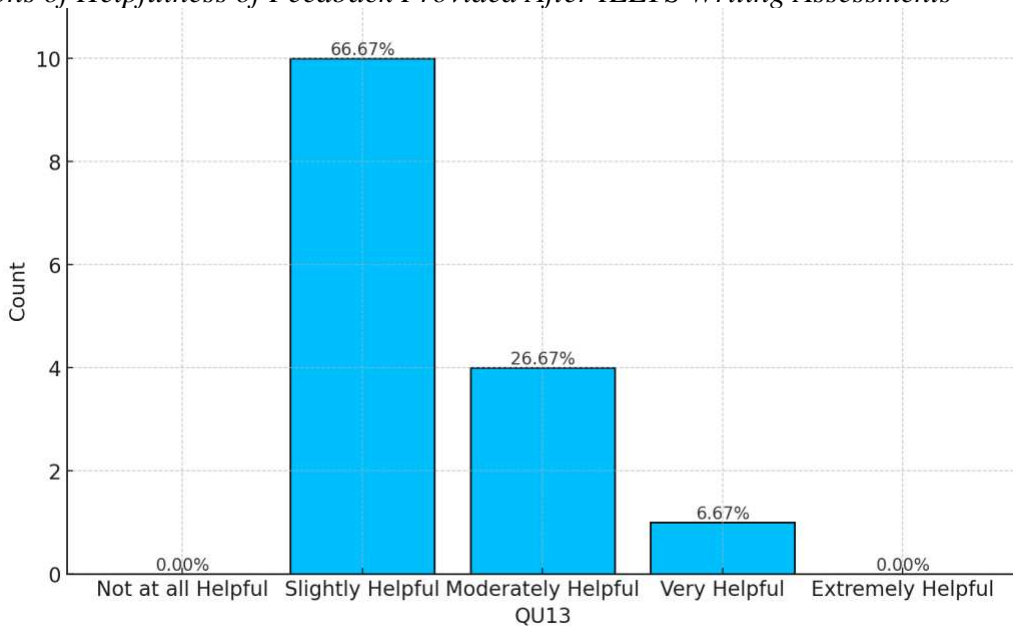


As shown in Figure 8, 60% of test-takers slightly understand scoring criteria used in IELTS writing assessment while 33.33% understand them moderately. 6.67% of respondents do not understand the criteria at all and none reported fully understanding them.

Q 13. Do you believe feedback provided after IELTS writing assessments is helpful for improving your writing skills?

Figure 9

Perceptions of Helpfulness of Feedback Provided After IELTS Writing Assessments

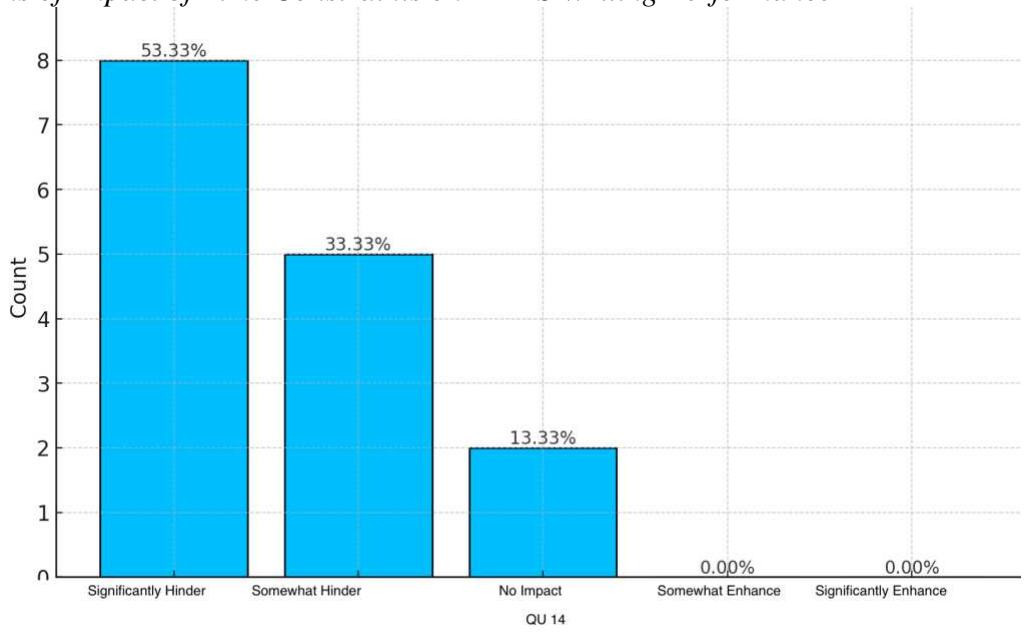


As illustrated in Figure 9, 66.67% of test-takers find feedback provided after IELTS writing assessments to be slightly helpful in improving their writing skills while 26.67% find it moderately helpful. Only 6.67% of respondents find feedback very helpful.

Q 14. How do you feel time constraints in the IELTS writing assessment affect your performance?

Figure 10

Perceptions of Impact of Time Constraints on IELTS Writing Performance

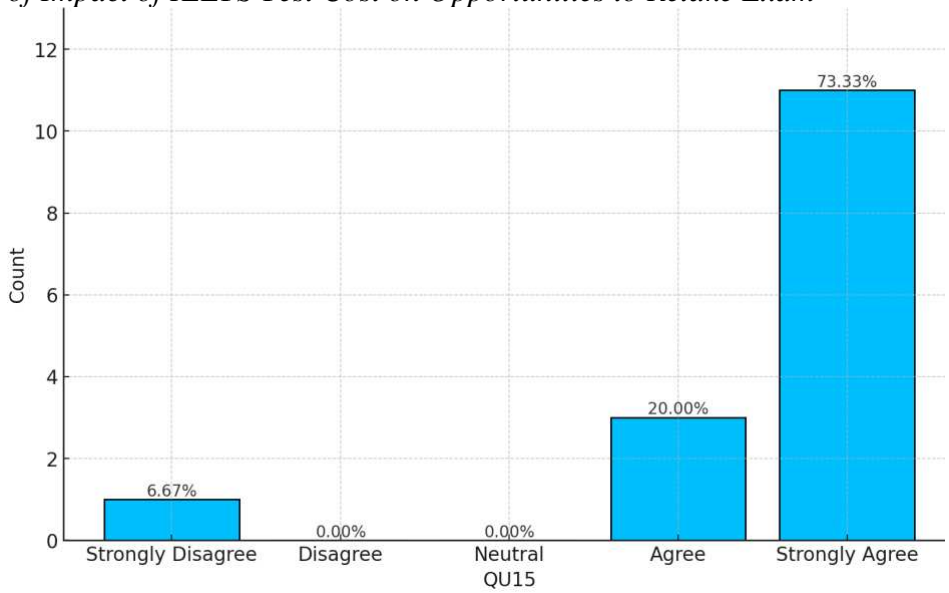


According to Figure 10, 53.33% of respondents believe that time constraints in IELTS writing assessment significantly hinder their performance while 33.33% feel constraints somewhat hinder their performance. Only 13.33% of respondents think time limits have no impact on their performance and none reported an improvement in performance due to time constraints.

Q 15. Do you feel that the cost of the IELTS test limits your opportunities to retake the exam?

Figure 11

Perceptions of Impact of IELTS Test Cost on Opportunities to Retake Exam

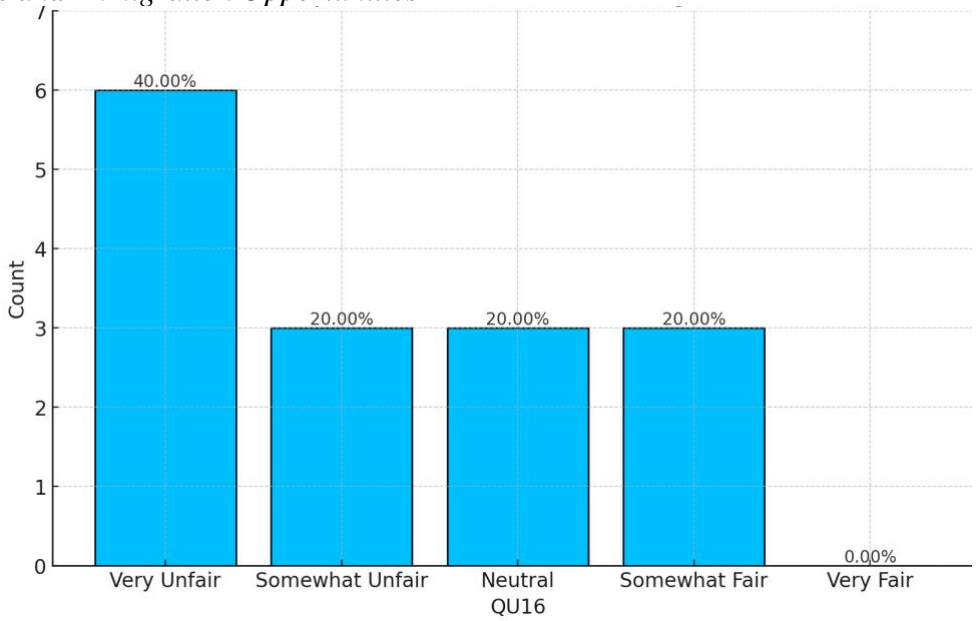


As indicated by Figure 11, 73.33% of respondents strongly agree that cost of IELTS test limits their opportunities to retake the exam while 20% agree with this statement.

Q 16. How fair do you find it that your IELTS writing score could impact university admissions, job placements or immigration opportunities?

Figure 12

Perceptions of Fairness in Impact of IELTS Writing Scores on University Admissions, Job Placements and Immigration Opportunities



According to Figure 12, 48% of respondents believe it is very unfair that the IELTS writing score could impact university admissions, job placements, or immigration opportunities, while 20% feel it is somewhat unfair. 20% of respondents feel neutral and an equal 20% consider impact somewhat fair.

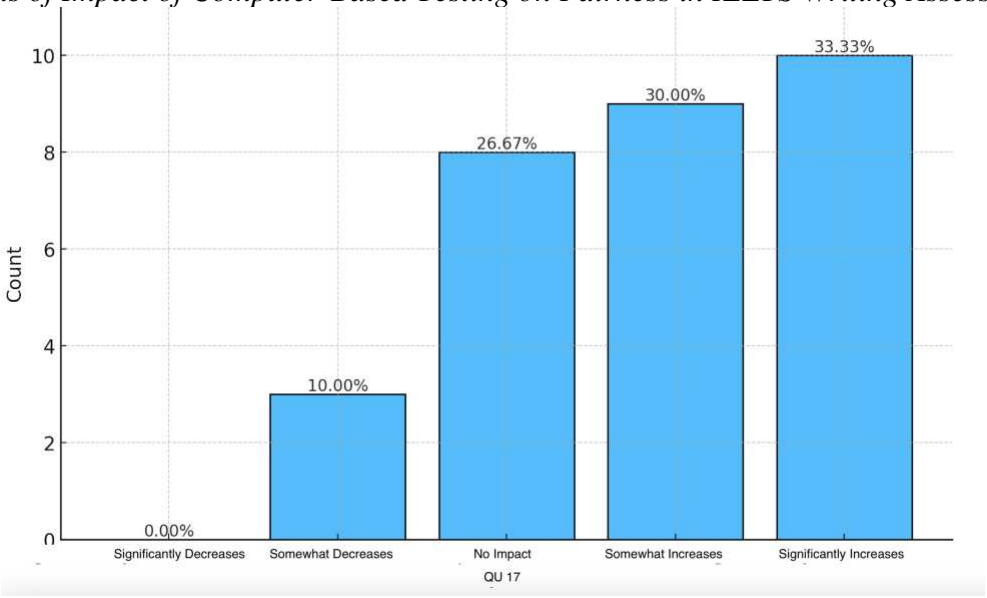
Section C: Impact of Technology on Fairness in IELTS Writing Assessment

(Questions 17 to 20)

Q 17. How do you feel use of a computer-based test (e.g., typing your essay) affects fairness in IELTS writing assessment?

Figure 13

Perceptions of Impact of Computer-Based Testing on Fairness in IELTS Writing Assessment

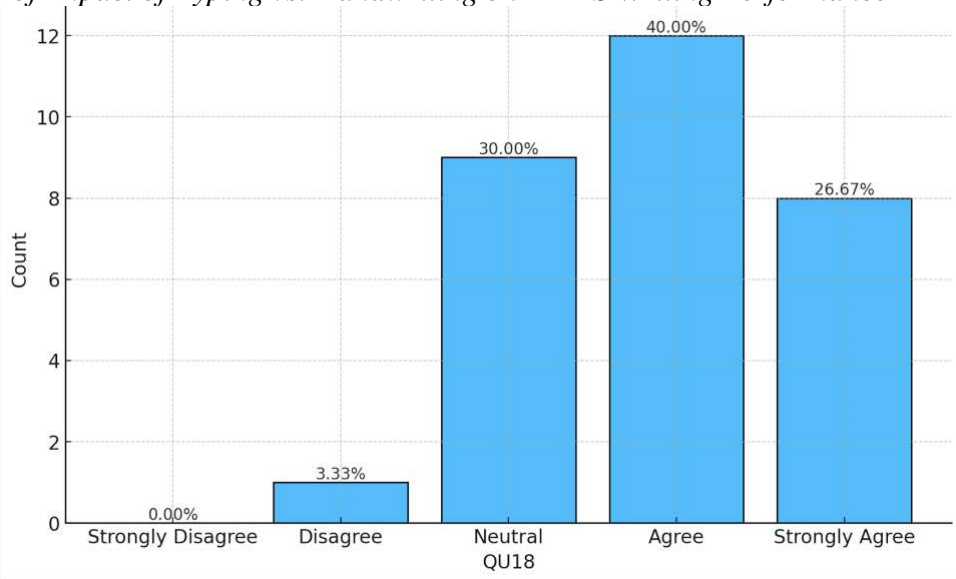


As shown in Figure 13, 33.33% of respondents believe that using a computer-based test significantly increases fairness in IELTS writing assessment while 30% think it somewhat increases fairness. A smaller portion (26.67%) feel that using technology has no impact on fairness and only 10% believe it somewhat decreases fairness.

Q 18. Do you think typing your essay instead of handwriting it affects your performance in IELTS writing task?

Figure 14

Perceptions of Impact of Typing vs. Handwriting on IELTS Writing Performance

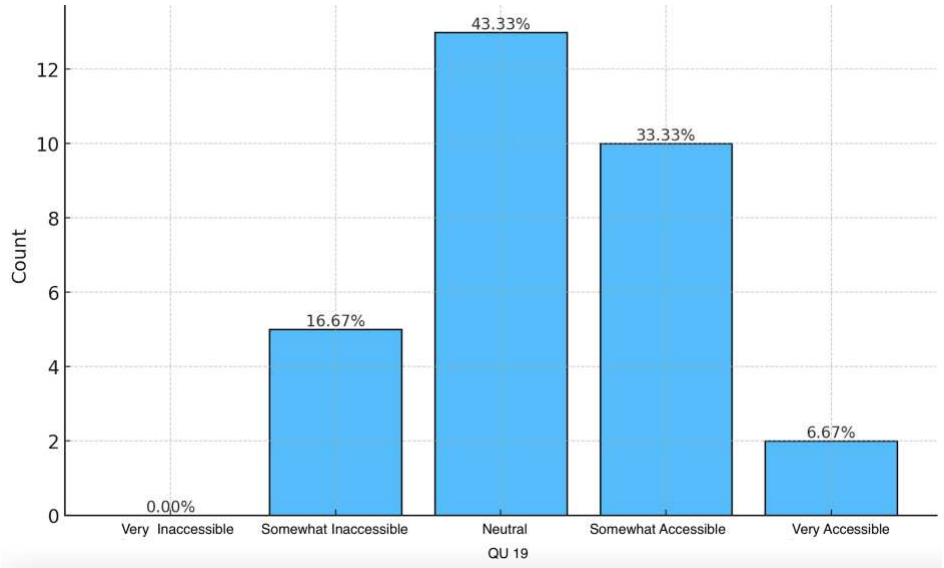


As illustrated in Figure 14, 26.67% of respondents strongly agree that typing their essay instead of handwriting it affects their performance in the IELTS writing task while 40% agree with this statement. Only 3.33% of respondents disagree and 30% remain neutral on impact of typing versus handwriting.

Q 19. How accessible do you find online resources and practice tools for IELTS writing preparation?

Figure 15

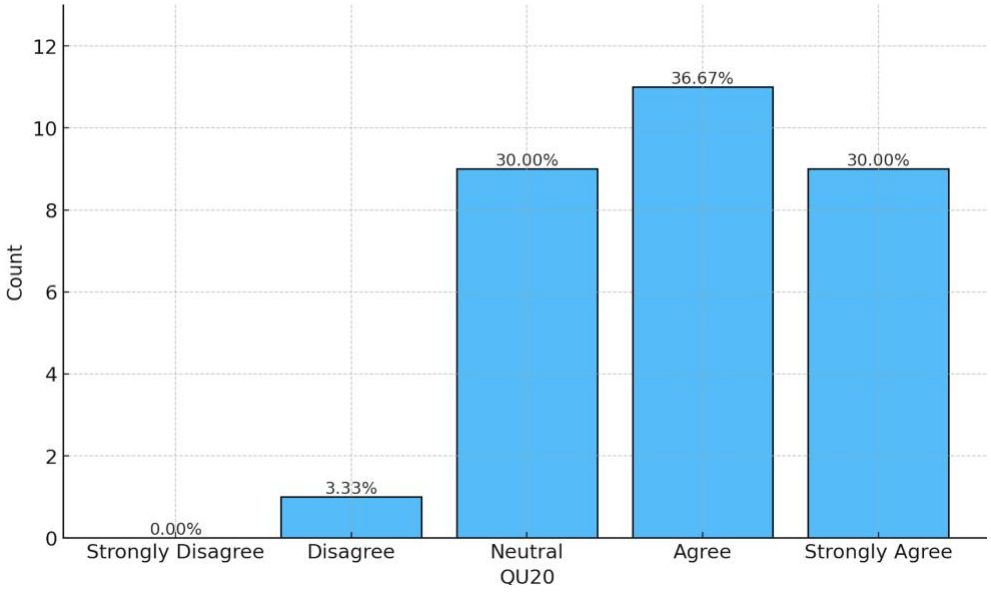
Perceptions of Accessibility of Online Resources and Practice Tools for IELTS Writing Preparation



As depicted in Figure 15, 33.33% of respondents find online resources and practice tools for IELTS writing preparation somewhat accessible while 43.33% hold neutral opinions. 16.67% of respondents find these resources somewhat inaccessible and 6.67% consider them very accessible.

Q 20. Do you believe that using technology (such as online test preparation, automated feedback) has improved your IELTS writing performance?

Figure 16
Perceptions of Impact of Technology on IELTS Writing Performance



According to Figure 16, 30% of respondents strongly agree that using technology such as online test preparation and automated feedback has improved their IELTS writing performance while 35.67% agree. Only 3.33% disagree and 30% hold neutral opinions on impact of technology on their writing performance.

Then, Wilcoxon Signed-Rank Test and t-tests were employed to compare observed responses to their respective means and determine whether the responses significantly deviated from the average. These tests are selected according to the nature of data and number of participants. Wilcoxon Signed-Rank Test is among the non-parametric statistical tests that

compare paired data when distribution of differences between pairs cannot be assumed normal. This test was appropriate because data collected from participants were ordinal and subjective and represented perceptions and opinions that might not follow a normal distribution. Additionally, sample size of 15 participants for both Section A (educators/examiners) and Section B (test-takers) was relatively small which justifies the use of a non-parametric test like Wilcoxon Signed-Rank Test. For Section C, sample size was larger (N=30) and a t-test was used. T-test is appropriate for comparing means when data are normally or approximately normally distributed which allows us to see if an observed mean is significantly different from expected value. Using both non-parametric and parametric methods ensures that each section's statistical properties were accounted for, and offered accurate information perceptions about fairness in IELTS writing assessment.

Descriptive statistics summarize key characteristics of data including mean, standard deviation, minimum, maximum, and frequency distributions. The results below provide an overview of educators/examiners and test takers responses to selected elements of the IELTS writing assessment.

Table 2

Descriptive Statistics

Category	N	Minimum	Maximum	Mean	Median	Std. Deviation	P-Value
Educators and Examiners							
QU5	15	1	4	2.13	2.00	0.915	0.008
QU6	15	1	4	2.47	2.00	0.915	0.046
QU7	15	1	3	1.93	2.00	0.704	0.001
QU8	15	2	4	2.93	3.00	0.829	0.101
QU9	15	2	5	3.67	4.00	0.900	0.020
QU10	15	1	4	2.13	2.00	0.990	0.010
Test-Takers							
QU11	15	1	3	1.67	1.00	0.816	0.001
QU12	15	1	3	2.27	2.00	0.594	0.002
QU13	15	2	4	2.40	2.00	0.632	0.007
QU14	15	1	3	1.60	1.00	0.737	0.001
QU15	15	1	5	4.53	5.00	1.060	0.002
QU16	15	1	4	2.20	2.00	1.207	0.022
ALL							
QU17	30	2	5	3.87	4.00	1.008	0.000
QU18	30	2	5	3.90	4.00	0.845	0.000
QU19	30	2	5	3.30	3.00	0.837	0.060
QU20	30	2	5	3.93	4.00	0.868	0.000

The comparison was made using Wilcoxon Signed-Rank Test with each question checked against a fixed median value of 3 (in a 5-point Likert scale). This decision was based on non-parametric nature of data because the normality assumption was not met. The null hypothesis assumes that responses would equal this neutral median (3) while alternative hypothesis proposed that responses would differ. The median value of 3 was selected because it represents neutral midpoint of a Likert scale which is often interpreted as a 'neutral' or 'undecided' response in social science research (Boone & Boone, 2012; Wakita, Ueshima, & Noguchi, 2012). So, if test-takers did not perceive significant biases or problems in IELTS writing assessment, their responses would naturally center around this neutral point. Therefore, the neutral value was used as a reference to check whether participants' responses deviated significantly from it which shows whether their

views on fairness in IELTS writing assessment were neutral or skewed. In addition to reporting p-values, mean values are also included in the tables for additional clarity.

The results show that educators and examiners did not find the IELTS writing assessment criteria clear (mean = 2.13). They also believed that the IELTS writing assessment is not fair for test-takers from different cultural and linguistic backgrounds (mean = 2.47). The assessment addresses cultural differences only to a small extent (mean = 1.97). Test-takers, in turn, found it unfair that their IELTS writing score could impact university admissions, job opportunities, or immigration (mean = 2.2). This result underscores the social consequences of relying heavily on these scores in life-changing opportunities. The participants expressed concerns that the writing score disproportionately affects individuals' access to these opportunities, which they perceive as socially unjust. The data reveals that fairness in the IELTS test is generally considered low based on educators' and examiners' opinions. Test-takers also face several challenges that undermine fairness, such as insufficient time, high exam costs, and unclear instructions. However, technology and computer-based IELTS assessments are seen as improving fairness, as indicated by the scores for the technology-related questions (all above the mean).

Table 3

Results of T-tests and Wilcoxon Signed-Rank Test

Section	N	Mean	t or z	p-value
A	15	15.33	-5.87	<0.001
B	15	14.66	-2.67	0.008
C	30	15	5.42	<0.001

For sections A and C with normal data, t-test was used; for section B with non-normal data, Wilcoxon Signed-Rank test was applied. The results of Wilcoxon Signed-Rank Tests and t-tests indicated that the mean score for section A is 15.33, with a p-value of less than 0.05, indicating that this set of questions significantly differs from the midpoint, suggesting that educators and

examiners perceive the IELTS writing assessment criteria as unfair. Mean score for section B is 14.66 with a p-value of less than 0.05 indicating that test-takers face significant challenges in the IELTS writing assessment, such as insufficient time, high exam costs, and unclear instructions. The mean score for section C is 15, with a p-value of less than 0.05, showing that respondents believe that technology and computer-based IELTS assessments improve fairness.

Qualitative Analysis

Qualitative data which were obtained from interviews provided rich insights about perceptions of fairness in IELTS writing assessment. Using thematic analysis, four key themes were derived.

Theme 1: Unclear Scoring Criteria

One of the most prominent themes to emerge from the interviews was the lack of clarity and subjectivity in the scoring criteria, particularly with regard to Task Response. This issue extended to other areas of assessment but was most concerning in Task Response where participants showed significant inconsistencies and confusion. The Task Response criterion was consistently cited as highly subjective. Examiners and educators admitted that their interpretation of what constitutes a fully addressed task often differed, which resulted in inconsistent scoring across different examiners.

An educator explained:

"Task Response is hardest part to explain. We have these guidelines but they're not very concrete and clear. You could have two examiners looking at same essay and come to different conclusions about whether it fully addresses task."

An examiner shared a similar concern:

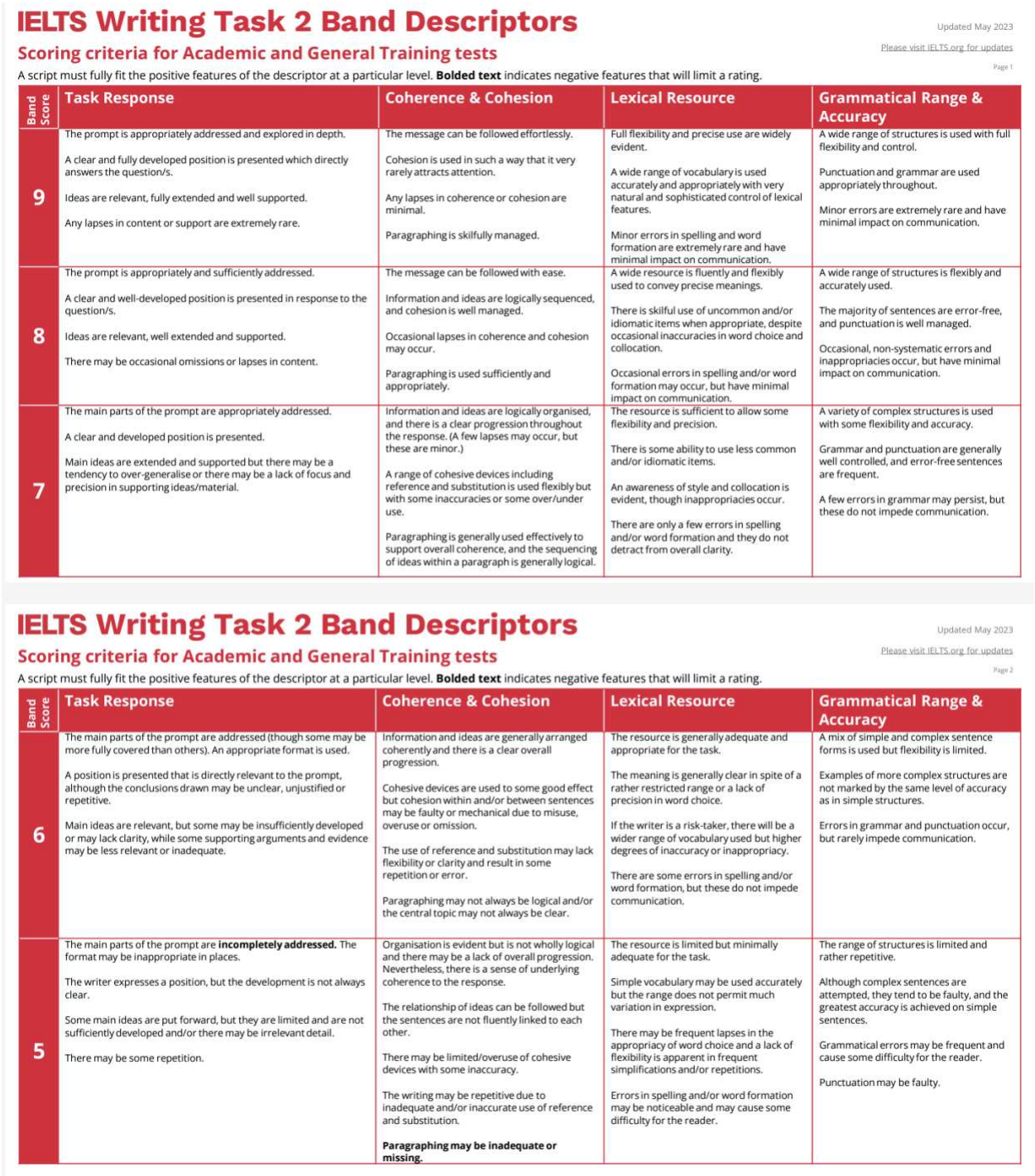
“Even with the guidelines, Task Response is open to interpretation. Some of my colleagues focus on addressing all parts of the task, while others are more concerned with the quality of the argument. It’s hard to be consistent.”

For test-takers, this inconsistency was incredibly frustrating. One test-taker mentioned: "I've taken the test three times and each time I feel like I've answered the question very completely. But my Task Response score never goes up and I don't know why. It feels like a guessing game."

The Task Response criterion lacks clear and specific benchmarks, which makes it difficult for both educators and test-takers to know what is required for higher scores. While Grammatical Range & Accuracy or Lexical Resource are more objective, Task Response descriptors are abstract, with terms like “appropriate” or “sufficiently addressed,” leaving much open to interpretation (see Figure 17).

Figure 17

Screenshot of the IELTS writing band descriptors PDF (British Council, n.d.).



IELTS Writing Task 2 Band Descriptors

Scoring criteria for Academic and General Training tests

Updated May 2023

Please visit IELTS.org for updates

Page 1

A script must fully fit the positive features of the descriptor at a particular level. **Bolded text** indicates negative features that will limit a rating.

Band Score	Task Response	Coherence & Cohesion	Lexical Resource	Grammatical Range & Accuracy
9	The prompt is appropriately addressed and explored in depth. A clear and fully developed position is presented which directly answers the question/s. Ideas are relevant, fully extended and well supported. Any lapses in content or support are extremely rare.	The message can be followed effortlessly. Cohesion is used in such a way that it very rarely attracts attention. Any lapses in coherence or cohesion are minimal. Paragraphing is skilfully managed.	Full flexibility and precise use are widely evident. A wide range of vocabulary is used accurately and appropriately with very natural and sophisticated control of lexical features. Minor errors in spelling and word formation are extremely rare and have minimal impact on communication.	A wide range of structures is used with full flexibility and control. Punctuation and grammar are used appropriately throughout. Minor errors are extremely rare and have minimal impact on communication.
8	The prompt is appropriately and sufficiently addressed. A clear and well-developed position is presented in response to the question/s. Ideas are relevant, well extended and supported. There may be occasional omissions or lapses in content.	The message can be followed with ease. Information and ideas are logically sequenced, and cohesion is well managed. Occasional lapses in coherence and cohesion may occur. Paragraphing is used sufficiently and appropriately.	A wide resource is fluently and flexibly used to convey precise meanings. There is skilful use of uncommon and/or idiomatic items when appropriate, despite occasional inaccuracies in word choice and collocation. Occasional errors in spelling and/or word formation may occur, but have minimal impact on communication.	A wide range of structures is flexibly and accurately used. The majority of sentences are error-free, and punctuation is well managed. Occasional, non-systematic errors and inappropriacies occur, but have minimal impact on communication.
7	The main parts of the prompt are appropriately addressed. A clear and developed position is presented. Main ideas are extended and supported but there may be a tendency to over-generalise or there may be a lack of focus and precision in supporting ideas/material.	Information and ideas are logically organised, and there is a clear progression throughout the response. (A few lapses may occur, but these are minor.) A range of cohesive devices including reference and substitution is used flexibly but with some inaccuracies or some over/under use. Paragraphing is generally used effectively to support overall coherence, and the sequencing of ideas within a paragraph is generally logical.	The resource is sufficient to allow some flexibility and precision. There is some ability to use less common and/or idiomatic items. An awareness of style and collocation is evident, though inappropriacies occur. There are only a few errors in spelling and/or word formation and they do not detract from overall clarity.	A variety of complex structures is used with some flexibility and accuracy. Grammar and punctuation are generally well controlled, and error-free sentences are frequent. A few errors in grammar may persist, but these do not impede communication.

IELTS Writing Task 2 Band Descriptors

Scoring criteria for Academic and General Training tests

Updated May 2023

Please visit IELTS.org for updates

Page 2

A script must fully fit the positive features of the descriptor at a particular level. **Bolded text** indicates negative features that will limit a rating.

Band Score	Task Response	Coherence & Cohesion	Lexical Resource	Grammatical Range & Accuracy
6	The main parts of the prompt are addressed (though some may be more fully covered than others). An appropriate format is used. A position is presented that is directly relevant to the prompt, although the conclusions drawn may be unclear, unjustified or repetitive. Main ideas are relevant, but some may be insufficiently developed or may lack clarity, while some supporting arguments and evidence may be less relevant or inadequate.	Information and ideas are generally arranged coherently and there is a clear overall progression. Cohesive devices are used to some good effect but cohesion within and/or between sentences may be faulty or mechanical due to misuse, overuse or omission. The use of reference and substitution may lack flexibility or clarity and result in some repetition or error. Paragraphing may not always be logical and/or the central topic may not always be clear.	The resource is generally adequate and appropriate for the task. The meaning is generally clear in spite of a rather restricted range or a lack of precision in word choice. If the writer is a risk-taker, there will be a wider range of vocabulary used but higher degrees of inaccuracy or inappropriacy. There are some errors in spelling and/or word formation, but these do not impede communication.	A mix of simple and complex sentence forms is used but flexibility is limited. Examples of more complex structures are not marked by the same level of accuracy as in simple structures. Errors in grammar and punctuation occur, but rarely impede communication.
5	The main parts of the prompt are **incompletely addressed.** The format may be inappropriate in places. The writer expresses a position, but the development is not always clear. Some main ideas are put forward, but they are limited and are not sufficiently developed and/or there may be irrelevant detail. There may be some repetition.	Organisation is evident but is not wholly logical and there may be a lack of overall progression. Nevertheless, there is a sense of underlying coherence to the response. The relationship of ideas can be followed but the sentences are not fluently linked to each other. There may be limited/overuse of cohesive devices with some inaccuracy. The writing may be repetitive due to inadequate and/or inaccurate use of reference and substitution. **Paragraphing may be inadequate or missing.**	The resource is limited but minimally adequate for the task. Simple vocabulary may be used accurately but the range does not permit much variation in expression. There may be frequent lapses in the appropriacy of word choice and a lack of flexibility is apparent in frequent simplifications and/or repetitions. Errors in spelling and/or word formation may be noticeable and may cause some difficulty for the reader.	The range of structures is limited and rather repetitive. Although complex sentences are attempted, they tend to be faulty, and the greatest accuracy is achieved on simple sentences. Grammatical errors may be frequent and cause some difficulty for the reader. Punctuation may be faulty.

One educator explained:

“With grammar or vocabulary, I can clearly tell my students what they need to improve, more complex sentences, fewer spelling errors, more idiomatic expressions. But with Task Response, it’s vague. What does it mean to ‘fully address the task’? It’s much harder to teach.”

A test-taker echoed this frustration:

"I know how to improve my grammar and vocabulary, but Task Response feels like a moving target. One examiner says I didn’t develop my ideas enough, but another says my ideas were clear. It’s impossible to know what’s expected."

There was also confusion about how Task Response overlaps with other scoring criteria, such as Coherence & Cohesion and Lexical Resource. Some aspects of the response, such as paragraphing or choice of vocabulary, are judged under multiple criteria, which creates confusion about where marks are being lost.

An examiner explained:

“Sometimes I find myself docking marks under both Task Response and Coherence & Cohesion for the same issue, like poor paragraphing or lack of logical flow. It’s difficult to separate these criteria because they influence each other.”

A test-taker added:

“I got feedback that my Task Response was weak because my paragraph weren’t organized well but isn’t it a part of Coherence and Cohesion? I am confused because I don’t know which area I should improve.”

Test-takers also mentioned receiving vague feedback regarding Task Response, which hindered their ability to improve their scores. While feedback on Grammatical Range & Accuracy

or Lexical Resource could be specific (for example, 'too many subject-verb agreement errors'), feedback on Task Response was often generic, such as 'did not fully address the task.'

One test-taker expressed their frustration:

“The feedback I got just said I didn’t fully address the task. But I don’t know what means. I followed the question exactly, so what else was I supposed to do?”

Another test-taker commented:

“I’ve gotten 6.5 for Task Response every time, but no one can tell me exactly why. They say things like ‘not an enough development’ or ‘didn’t address all parts,’ but it’s too vague and not helpful.” The Task Response criterion can vary significantly depending on the type of task (e.g., opinion essays, problem-solution essays, or discussion essays). The general rubric does not provide clear enough guidance for how to address different requirements of each task type which further complicates understanding.

One educator mentioned:

“Task Response changes depending on whether it’s a discussion or an opinion essay and that makes it harder for students. One of my students might do really well on one type but struggle with another because the requirements aren’t clear across different task types.”

A test-taker explained:

“I did well in opinion essay but when I got a problem-solution task I didn’t know how to structure my response properly. It’s not clear how much focus have to put on identifying problem versus solutions.”

Many test-takers raised concerns about how their overall IELTS writing score was affected by Task Response even though they performed well in other areas like Grammar and Lexical Resource.

As one test-taker reported:

"Generally, I always did well in grammar and vocabulary, but my overall score won't go above 6.5 because of Task Response. It's frustrating because I know I'm capable for better score but this criterion is holding me back. That is not fair something as subjective as Task Response can influence my overall grade so much. I'm good at grammar and vocabulary and even at text organization but it's not enough to get score I need."

Subjectivity of Task Response leads to concerns about fairness especially in high-stakes applications like university admissions, job placements or immigration decisions. Candidates may lose out on life-changing opportunities due to how one or two examiners interpret their Task Response.

One test-taker shared:

“I needed a 7 to get into my university program, but I missed it because of Task Response. I did well in grammar and vocabulary, but my overall score stayed at 6.5. It’s unfair because I know I’m qualified, but this one part of the test is stopping me.”

An educator added:

“I’ve had students who are brilliant in every other area, but because they didn’t meet the expectations for Task Response, they couldn’t get the score they needed for immigration. It feels like the system is stacked against them.”

As illustrated in the attached image (Figure 17), Task Response requires candidates to appropriately address and explore the prompt, with expectations varying based on the score band. However, these expectations are not as concrete as the criteria for Grammatical Range & Accuracy or Lexical Resource. For example, for Band 9, candidates must provide a "clear and fully developed position" and support their ideas with relevant details. For Band 7, candidates should

address main parts of prompt but may lack precision in supporting ideas. While Grammatical Range & Accuracy can be measured objectively (e.g., by counting errors), Task Response involves a level of judgment that leads to variation in interpretation and scoring.

As one examiner explained:

"Task Response is more open to personal interpretation. While I can count grammatical errors, I have to make a judgment call about whether response fully addresses task, and that's where things get tricky. This inconsistency with regard to the scoring of Task Response, for instance, as compared with more objective criteria such as Grammar or Lexical Resource, poses huge problems to candidates and instructors alike by casting questions in terms of fairness".

Theme 2: Cultural Bias in Writing Prompts and Scoring Practices

Perceived cultural bias in both the writing topics and scoring practices is one of the emerging themes from the interviews. Both test-takers and educators reportedly suspect that certain prompts favored Western perspectives at the expense of those coming from non-Western backgrounds. The scoring process also appears to have its own set of biases insofar as examiners coming from a Western background may unconsciously favor writing styles or ideas more familiar to them.

A specific example that one test-taker mentioned was a question that asked candidates to remark on the strengths of individualism, which is engrained deeply in Western cultures but not as deeply in more collectivist societies such as those found in East Asia or the Middle East thus test-takers from these regions felt they could not present themselves as well.

A teacher took this point:

"Some of the topics are very culturally specific. My students from China or the Middle East struggle because they have less encountered these concepts in their everyday lives. It is not right that we expect to write an essay on aspects they have no experience with for example 'imposing a curfew for teenagers not to be allowed to be out of door at night'. They already are scared of being out at night because of lack of safety in many cities in some eastern countries. How can they write about such a thing".

The test-takers complained about the difficulty of writing on topics which are either irrelevant or even remote from their cultural experiences.

One test-taker shared:

"I had to write about a topic that's not common in my country. It was hard because I didn't have any personal experience or knowledge to draw from and I think that affected my score."

Beyond the problems inherent in prompts themselves, some were concerned about the subjective impact of Western scorers. Indeed, some educators and test-takers questioned the fact that the Western panel of scorers may unintentionally give higher marks to those arguments and styles of writing that stem from a more Western approach. For example, in some countries, people are taught to write in an indirect, "zig-zag" style, which contrasts with the more direct approach commonly preferred in Western countries. If the scorers are predominantly from Western backgrounds, they may be unconsciously biased towards direct writing styles, potentially affecting their scoring of non-Western test-takers.

As one examiner admitted,

"It's not deliberate, but there may be a leaning towards giving higher marks to the essays which have been presented as scorers are accustomed to seeing the presentation of arguments in

the West. That puts some candidates at a disadvantage because their writing style may reflect their cultural background".

Such a dual-layered bias in both prompts and scoring practices creates an unequal playing field for test-takers from outside the West thereby perpetuating a perception of unfairness in the IELTS writing assessment.

Theme 3: Life-Changing Consequences

One of the most important themes which was emotionally charged was the significant impact that low IELTS writing scores had on test-takers' lives. Many shared personal stories of how a single relatively low score in writing despite strong performances in other language skills had delayed or derailed their academic, career and immigration plans.

A particularly moving story came from a test-taker who missed the chance to apply to McGill University's prestigious engineering department. He explained:

"I took the IELTS three times, each time missing the required 6.5 in writing by half a point, despite scoring over 7 in all the other sections. I finally missed the McGill application deadline because of that writing score. After filing a complaint, my score was increased by half a point, and I got into McGill—but I had already lost an entire year of my life. It's devastating, knowing it could have been avoided."

In addition to academic barriers, the consequences of writing scores extended into immigration as well.

Another participant who had hoped to apply for permanent residency in Canada, shared the following:

"I needed at least 6.5 in writing for my PR application. I scored 7 and higher in all the other sections but only managed to get 6.0 in writing. That half-point difference meant I had to wait an

entire year to reapply. It wasn't just a test—it was my future, my chance to start a new life in Canada. I felt completely defeated."

The impact of these stories goes beyond test scores, deeply affecting people's lives. One test-taker who had worked hard to excel in listening, reading, and speaking said:

"I got an 8 in listening and a 7 in speaking, but my writing score was just 6.0. Because of that, I couldn't apply to my dream university in Canada. It feels like all my hard work was wasted."

The emotional toll is exacerbated by the perception that writing scores, often seen as subjective, are disproportionately weighted in critical life decisions. A participant who lost out on a job in the US shared:

"I got 6.0 in writing when I needed 6.5. I did well in every other section, but that half-point meant I lost the job. It doesn't feel fair at all."

These stories reveal the profound consequences that minor differences in writing scores can have on people's lives, delaying academic pursuits, job opportunities, and even immigration plans.

As one test-taker tearfully expressed:

"It's not just a test score; it's my life. Missing out on these opportunities because of half a point feels like the system is working against us. I am sure the examiners themselves are not certain about the scores they give to us because after filing complaints, many scores have changed. But all people cannot complain because it is expensive".

The emotional and social consequences of this system leave many feeling trapped, their futures on hold, due to narrow thresholds that may not fully reflect their true language abilities.

Two of participants said that a further analysis of their writing scores revealed that their low scores were due to difficulties with task response, rather than other criteria such as grammar,

vocabulary, or coherence and cohesion. A detailed review of a test-taker's essay, for instance, showed that while they had demonstrated strong grammar and advanced vocabulary, they had not developed the task prompt. The test-takers focused too much on discussing a specific aspect of the question, which led to a lower score in task response. A mistake, irrelevant to language skills, led to missing golden opportunities in their life.

Theme 4: Positive Impact of Technology on Fairness

Integration of technology within IELTS writing test is generally considered a step forward in enhancing fairness and equity. The test candidates reported generally favorable impressions of the computer-based testing (CBT), highlighting the ways in which this testing format reduced stress and offered a more accessible and more user-friendly venue for writing tests. They believed that it was much easier to type their essay on computer compared to handwriting. For instance, one test taker explained the following:

"I was used to typing and being able to type my essay really made much difference. I was not concerned with my handwriting; besides, it was easier to organize my thoughts. It felt like I had more control over the test."

This sentiment was echoed by some other test-takers especially those who found handwriting as a stressful component of paper-based exams. The ability to make quick edits without having to cross out mistakes or rewrite sentences was seen as a significant advantage. Another participant reflected:

"With the paper-based test, I was constantly worried about making mistakes because I couldn't erase them cleanly. On the computer, I could just delete or move things around. It made me less anxious and allowed me to focus more on my ideas."

Regarding fairness, test-takers argued that the computer-based format made it a more level playing field especially for those participants who struggled with the very act of writing or were challenged by a disability. Accessibility features related to computer-based testing, such as font size adjustment, screen readers, and spell checkers, were listed as enhancements that would serve the needs of a wide range of test-takers. One educator commented:

"For students with dyslexia and problems in motor skills, being able to type out their responses and use assistive technology all makes a difference. Technology has also changed the way test-takers prepare for IELTS writing section. Some of participants said during practice for the test, they use automated feedback to improve their writing by using features like grammar checkers, spell checkers and essay evaluation sites. These tools provided immediate feedback allowing test-takers to identify common errors and make improvements before the test.

One test-taker shared their experience with automated feedback:

"I used an online writing checker that gave me a score based on my grammar, vocabulary and organization. It was helpful because I could see where I was going wrong and fix it. It was like having a tutor without cost."

However, participants also acknowledged the limitations of these tools. While they were useful for addressing technical aspects of writing such as grammar and coherence, they were less effective in helping candidates understand deeper components of writing task such as how to respond to the prompt effectively. One test-taker said:

"The grammar checker helped me catch mistakes, but it didn't explain how to improve my task response. I still struggled with understanding what examiners were looking for in terms of content."

Despite these few limitations, the majority of participants felt that technology provided a

useful supplement to traditional study methods, enabling them to practice their writing skills more efficiently. One test-taker noted:

"Before I used these tools, I wasn't sure where I was going wrong. The automated feedback helped me focus on my weak areas, and over time, I noticed an improvement. It gave me more confidence going into the test."

Technology also played a crucial role in improving fairness of scoring process. Some participants expressed the belief that use of automated scoring systems, although not currently implemented in IELTS, could potentially reduce human bias and ensure more consistent evaluations. One participant said:

"Human examiners have different opinions, and I've heard stories of people getting very different scores after a re-mark. If a computer could do part of the scoring, it might be fairer because it would focus on actual writing instead of being influenced by examiner's preferences."

While human oversight remains necessary for evaluating more subjective elements of writing, participants felt that technology could help mitigate instances of unfair scoring by providing an additional layer of consistency. An educator agreed, stating:

"Technology can be a great tool for standardizing grading process. It removes some of personal biases that human examiners might bring to the table which can make a real difference for candidates from different cultural backgrounds."

Some participants raised concerns about the potential for technology to create new inequalities. Not everyone has equal access to digital resources or is equally familiar with using computers which could disadvantage candidates from less technologically developed regions. One test-taker expressed this concern:

"In my country, there are only three cities where you can take the computer-based IELTS. For people living in rural areas or other cities it's much harder to access the test. It's great for people like me who use technology all the time, but I can see how it might be unfair to others who have to travel long distances or aren't comfortable using computers."

Some educators emphasized the need for test administrators to ensure that all candidates, regardless of their technological proficiency or location, have access to resources and training to prepare for computer-based tests. One educator suggested:

"We need to make sure that candidates from all backgrounds have access to the necessary training and technology before they take the test. Otherwise, the technology that is meant to improve fairness could end up creating new barriers."

Addressing the Research Questions with Qualitative Data

Research Question 1: How do educators and examiners perceive the fairness of the IELTS writing assessment criteria?

Educators and examiners have shown their concerns regarding fairness of IELTS writing assessment, in relation to clarity of scoring criteria particularly with respect to the Task Response criterion which seemed subjective and poorly defined. This vagueness of the scoring criteria of IELTS Writing, especially about the Task Response, led to a lack of consistency in how the examiners apply the criteria. Inconsistency meant that different examiners sometimes awarded varying scores on similar essays.

Also, with the partial overlaps of Task Response with other criteria, like Coherence & Cohesion, the scoring also became more complex and challenging; educators could give less clear guidance to their students. Unlike more quantifiable criteria like Grammar, Task Response had no clear-cut benchmarks, thus making things more complicated for examiners and educators alike when having to assess or teach it.

In summary, educators and examiners believed that subjectivity in the scoring process, the inconsistent application of the criteria, and ambiguity in the assessment guidelines all threatened the fairness of the writing assessment.

Research Question 2: What challenges do test-takers face in IELTS writing assessments in terms of fairness?

Test-takers faced a number of challenges regarding the fairness of the IELTS writing assessment. Among them, vagueness in the Task Response criterion was the biggest challenge. In the criterion of Task Response, many candidates felt that the feedback given was inadequate and too general to understand how the scoring pattern could be improved.

Another serious challenge was that low writing scores caused lifetime consequences for the candidates. Some test-takers narrated how a single score in writing diminished their chances of admission to university, job offers, or immigration applications. So much subjectivity in assessing the writing performance, especially in Task Response, made test-takers question how their overall scores were far too disproportionately affected by one criterion when scores in other areas like Grammar and Lexical Resource were good.

Besides this were significant financial costs of repeated IELTS re-takings or filing complaints in hopes of improving one's writing score, and importantly, timing constraints within

the actual test that made it impossible for candidates to work through a full elaboration of their response.

Taken altogether, test-takers faced challenges linked to subjectivity within the Task Response criterion, emotional and social consequences because of low scores, and time and financial pressures of re-taking the exam.

Research Question 3: How do technological tools and resources influence the fairness of scoring in IELTS writing assessments?

On the whole, technology had a positive impact on the fairness of IELTS writing assessment. Most test-takers viewed the advent of Computer-Based Testing as a positive development, as it is more convenient and less stressful than the paper-and-pencil test. They welcomed typing essays and the ease of editing without the headaches brought about by illegibility concerns that can be very anxiety-provoking during the test.

The automated feedback provided through tools used in test preparations was very helpful in grammar and coherence. These tools consistently provided useful feedback to test-takers right away for systematic identification and addressing weaknesses; however, the tool was less helpful in helping test-takers improve in the Task Response, which remained a subjective criterion that requires human judgment.

Technology helped reduce human bias and made the test more accessible, though it also raised concerns about digital inequality. In this regard, not all examinees had equal opportunities to participate in regions where access to a computer-based IELTS exam or technological resources may be limited, and new barriers to fairness could be created. Technology thus assisted the IELTS

Writing assessment to achieve apparent gains in certain areas of fairness; however, this was not without its limitations.

Discussion

This section is a discussion about the results of the three research questions proposed in this study based on both qualitative and quantitative findings. The discussion will integrate results with relevant literature, comparing and contrasting findings with available studies on fairness.

Research Question 1: How do educators and examiners perceive the fairness of the IELTS writing assessment criteria?

Both educator and examiner findings indicate concerns regarding IELTS writing test fairness particularly for Task Response criterion. Qualitative interviews had identified that assessment criteria are not clear. In particular, Task Response as one of four major assessment factors was seen by both educators and examiners as the most subjective and inconsistent criterion during scoring. Examiners themselves recognize variation in the way they interpret this criterion, which influences candidate score outcomes. This was reinforced by quantitative data, where a big percentage of the respondents were dissatisfied with the clarity in scoring criteria.

The analyses may suggest a lack of concrete benchmarks and variations in scoring which is in line with prior research by Hamid et al (2019) as an empirical study, and what was mentioned by Weir (2005) and Kunnan (2004), that generally state the threat to fairness usually comes from lack of clarity, and variability in examiner judgment.

Overlapping between Task Response and other criteria, such as Coherence & Cohesion, further problematizes the issue of fairness in the assessment. The educators and examiners felt frustrated by difficulty in determining at what point marks were being lost-a key concern in high-stakes situations where accurate feedback is needed. In many cases, the overlap among some assessment criteria led to examinees being doubly penalized.

This difficulty with high-stakes assessments when it relates to overlapping criteria has also been documented in research studies conducted by Hamid et al. (2009). The findings indicate that fairness might be increased by providing a more elaborated exposition of the Task Response criterion. This would fall in line with Kunnan's Fairness Framework (2000) where need for transparency and consistency of a high-stakes test environment is underlined.

Research Question 2: What challenges do test-takers face in IELTS writing assessments in terms of fairness?

Based on the results of both quantitative and qualitative analyses, some serious fairness issues in the IELTS writing assessment have been pointed out by test-takers. The major one is the subjectivity of scoring factors, especially Task Response criterion. According to the current study's findings, a number of test-takers could not understand why the score for this aspect did not improve though they had taken several tests. This point of uncertainty was enhanced by the vague feedback they received which did not make their performance clear, and nor indicated how the examinee could further improve.

This corroborates some other studies which highlight that a lack of clarity in scoring criteria makes test takers unsure about how to improve because the feedback they receive is often too ambiguous to be helpful. This exactly matches the current study's findings, in which test takers

indicated surprise over why their Task Response scores did not improve despite several attempts and even tutoring. A lack of transparency in scoring which makes it difficult for applicants to comprehend criteria they must reach was a problem replicated in our findings where unclear examiner response exacerbated test-takers' doubt. Furthermore, subjectivity in scoring particularly for writing tasks leads to disagreement among examiners as indicated in both other research and our study.

The most striking theme obtained from the results was one related to life-changing consequences of scores in writing. Many participants reported that a relatively low score in writing had delayed or cancelled their plans for university admissions, job opportunities or immigration.

This finding aligns with Hamid et al. (2019) who documented social and emotional toll of low scores on high-stakes language assessments like IELTS. In their study, test-takers showed concerns about subjectivity of test scoring and how even small differences in scores could have significant, life-changing impacts. They also noted that test-takers often felt IELTS functioned more as a gate-keeping tool for immigration and education rather than an accurate measure of language ability. This resonates with unfairness perceived by participants in present study whose overall capabilities were not fully reflected in their writing scores.

In addition, the findings show the perception that test writing prompts were biased towards particular cultures. According to the respondents, some writing topics gave an unfair advantage to their Western counterparts. This issue of cultural bias has been widely documented in the literature and, as Arefsadr and Babaii (2023) illustrate, certain cultural norms embedded in the test prompts can place test-takers from non-Western cultures at a disadvantage.

Other issues to come out were additional costs associated with retakes and time constraints in the actual test, from both the quantitative and qualitative data. These findings are consistent with

studies like the one conducted by Hamid et al. (2019), who documented how repeated testing for minor score improvements exacerbated financial burdens on test-takers. This suggests that test administrators should reconsider how writing prompts are set and reduce the need for costly retakes and lowering the stress associated with time constraints.

Research Question 3: How do technological tools and resources influence the fairness of scoring in IELTS writing assessments?

The findings of quantitative and qualitative sections of the present study showed that an introduction of technology in the IELTS writing assessment was believed to bring a generally positive impact on fairness, through the use of computer-based testing. Computer-based testing was favored because it provided test candidates with easier ways to type essays and make corrections. Such benefits support the studies by Hamp-Lyons (2016), who cited that technology may reduce human error and improve test accessibility.

There were, however, concerns regarding the digital divide, especially for regions where access to the computer-based IELTS tests is at a minimum. It addresses the concern of Weir (2005) about ensuring equal access to testing resources for all candidates. Moreover, automated feedback tools were widely used by test-takers during their preparation, particularly for improving grammar and coherence. While these tools are helpful, they may be less effective in the Task Response criterion than other criteria, since this is a more subjective requirement calling for human judgment.

This supports Hamp-Lyons' (2016) observation that while equally effective as human raters in ranking mechanical aspects of the writing, the performance of automated systems lags in terms of deeper textual features such as argumentation and task response. However, inequities regarding

digitally disadvantaged learners need to be resolved properly. Future efforts must be directed towards widening access to CBT and enhancing the capacity of automated tools to address even more subjective criteria like Task Response.

Conclusion

This chapter has presented both quantitative and qualitative analyses that offer critical insights about perceptions of fairness in IELTS writing assessments. The quantitative findings validated by Cronbach's alpha and supported by expert review highlight concerns about subjectivity of scoring process, particularly with the Task Response criterion. This subjectivity has led to inconsistencies in scoring which educators and test-takers both find frustrating due to lack of clear benchmarks.

In qualitative analysis, one of the key themes was unclear and subjective nature of scoring criteria, particularly for Task Response, which often resulted in confusion among both test-takers and examiners. This ambiguity led to inconsistent scoring and made test-takers uncertain about how to improve their performance. This analysis also revealed themes of cultural bias in prompts and among scorers, significant life-changing consequences of low writing scores, and impact of technology on test fairness. While integration of technology was perceived as improving fairness through computer-based testing and automated feedback, concerns about unequal access to technology and limitations of automated tools in addressing subjective criteria like Task Response remain.

Chapter 5

Conclusion, Implications, Suggestions

Conclusion

The present study has explored issue of fairness in IELTS writing assessments. This research used both qualitative and quantitative methodologies to reveal a number of critical insights into educators', examiners' and test-takers' experiences, especially in scoring subjectivity and cultural bias and life-changing consequences of writing scores.

The current study has demonstrated that though IELTS is claimed to be one of the internationally accepted tools in assessing language proficiency, it is not without biases and inconsistencies that arise, particularly in high-stakes areas such as admissions to universities, job placements, and immigration prospects. Among the major emerging from this study is the fact that ambiguity and subjectivity of scoring factors, especially those related to the Task Response criterion, is a major reason for frustration among test-takers and inconsistency in the scores. This inconsistency leads to less fairness and has life-changing consequences for candidates, affecting their future opportunities.

Cultural bias as highlighted by both test-takers and educators remains a significant issue especially in writing prompts. The study suggests that culturally neutral prompts and examiner training in cultural awareness may help mitigate this bias. Additionally, clearer and more consistent scoring rubrics are essential to reduce subjectivity and variability in scoring, particularly in Task Response criterion.

This study identifies how technology can play a facilitative role in enhancing fairness through the use of computer-based testing. While the findings indicate that technology increases scoring consistency, especially in more mechanical aspects like grammar and coherence, it does not really challenge more deep-seated issues of subjectivity inherent in the Task Response criterion. Additionally, there are digital divide issues which emphasize equal access to technology, as unequal access may exaggerate existing inequalities among test-takers.

Implications

The findings of this study suggest some main areas of IELTS writing assessment that can be improved to ensure greater fairness and reduce unintended negative consequences.

Firstly, scoring factors especially Task Response criterion should be clarified more to provide both examiners and test-takers with more concrete guidelines. This would reduce subjectivity that currently undermine fairness of assessment. Clearer guidelines would help ensure that test-takers understand how to fully address the task thereby improving their chances of meeting the required writing scores. Examiners would also benefit from clearer rubrics which result in more consistent evaluations across different test centers.

Secondly, getting poor writing scores has important emotional and social consequences. Many test takers claimed life-changing impacts such as delayed university admissions, missed job opportunities and lengthy wait times for immigration applications. These cases increase pressure on test-takers especially when a small gap in writing score prevents them from achieving their overall goals. Test developers and policymakers should consider reducing the emphasis placed on

a single writing score in high-stakes decisions and instead focus on more holistic evaluations of a candidate's language proficiency.

Thirdly, cultural bias should be addressed. This kind of bias emerged as a critical issue, both in the writing prompts and in the scoring process. Test designers should ensure that prompts are culturally neutral and do not favor candidates from specific backgrounds, thus providing an equal opportunity for all test-takers. This aligns with Kunnan's Fairness Framework (2000), which emphasizes the importance of cultural sensitivity in test design.

In addition, potential cultural differences in writing styles between test-takers and examiners should be considered. Some test-takers expressed concerns that their writing styles, which may reflect the norms of their cultural backgrounds, might not always align with the expectations of examiners who are more familiar with Western approaches to argumentation and expression. Although there is no definitive evidence of bias, raising examiners' awareness of cultural diversity in writing could help ensure that evaluations are based on language proficiency rather than differences in stylistic conventions.

Finally, efforts should be made to expand access to computer-based testing, particularly in regions where access is limited. Ensuring that all candidates have equal access to CBT will help reduce inequalities created by the digital divide. Additionally, CBT offers advantages such as greater consistency in scoring and reducing handwriting-related bias. Moreover, the presence of widely accepted home edition tests further increases accessibility, allowing candidates from remote or underserved areas to take the test without the need for a testing center, thus fostering greater fairness in the testing process.

By addressing these areas, IELTS writing assessment can become fairer, culturally sensitive and aligned with the diverse needs of global test-takers.

Limitations of the Study

While this study included people from various countries, the sample size of interview participants was rather limited which makes it difficult to generalize findings to all IELTS test takers. Another limitation is reliance on self-reported data, particularly in qualitative interviews which might be influenced by participants' subjective experiences. Furthermore, this study did not account for differences in educational backgrounds or access to test preparation tools and it may have influenced outcomes.

Suggestions for Further Research

Future research should look into the clarity of scoring criteria not only in IELTS, but also in other high-stakes language tests like TOEFL and Cambridge English exams, to see if similar issues of subjectivity and inconsistency develop. As this study has shown, the ambiguity surrounding Task Response poses substantial issues for both test takers and examiners, indicating the need for more explicit guidelines across various tests. Comparative research of different language proficiency exams could provide a more comprehensive knowledge of how scoring differences affect test outcomes and fairness on a worldwide scale. Furthermore, an investigation of cultural bias in writing prompts across multiple exams could shed light on whether these biases disadvantage non-Western applicants, as seen in our study.

Another area for future research would be to explore how new technologies, which might include automatic scoring systems and AI-based tests, are used to decrease subjectivity and bias in testing students' writing. These studies may look into the reliability and precision of machine-

generated ratings compared to human raters to see if such technologies can eliminate subjective problems that human examiners experience, particularly with regard to Task Response.

Finally, there is a dire need for more studies on developing mechanisms to improve feedback for test takers. More practical detailed feedback systems are great ways to help candidates recognize their weak points and improve fairness.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1985). T*he Standards for Educational and Psychological Testing.* Washington, DC: American Psychological association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). T*he Standards for Educational and Psychological Testing.* Washington, DC: American Psychological association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). T*he Standards for Educational and Psychological Testing.* Washington, DC: American Psychological association.

Arefsadr, A., & Babaii, E. (2023). Let their voices be heard: IELTS candidates' problems with the IELTS academic writing test. *TESL-EJ, 27(1).* https://doi.org/10.55539/10

Azizi, Z. (2022). Fairness in assessment practices in online education: Iranian University English teachers' perceptions. *Language Testing in Asia, 12*(1), Article 14. https://doi.org/10.1186/s40468-022-00164-7

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Barkaoui, K. (2016). *What changes and what doesn't? An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts*. IELTS Research Report Series, (3).

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension, 50*(2)

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

British Council. (n.d.). *IELTS around the world.* https://www.britishcouncil.org/exam/ielts/international

British Council. (n.d.). *IELTS writing band descriptors* [PDF]. https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

British Council. (2023). *Understanding IELTS.* https://www.britishcouncil.org/exam/ielts

Brown, J. D. (2000). *Using surveys in language programs*. Cambridge University Press.

Bryman, A., & Bell, E. (2011). *Business research methods* (3rd ed.). Oxford University Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Chapelle, C. A., & Douglas, D. (2006). *Assessing Language Through Computer Technology*. Cambridge University Press.

Erickson, G., & Tholin, J. (2022). Overall, a good test, but…: Swedish lower secondary teachers' perceptions and use of national test results of English. *Languages*, *7*(2), 73. https://doi.org/10.3390/languages7020073

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Fulcher, G. (2013). *Practical language testing*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Gokturk, A. L., & Tsagari, D. (2022). Evaluating perceptions towards the consequential validity of integrated language proficiency assessment. *Languages*, *7*(1), 65. https://doi.org/10.3390/languages7010065

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

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language testing in Asia, 9*(16), 1–20

Hamp-Lyons, L. (2001). Ethics, fairness(es), equity and the NNES issue. *Assessing Writing, 8*(2), 101-116.

Hamp-Lyons, L. (2016). The use of automated scoring in assessing writing. In Tsagari, D., & Banerjee, J. (Eds.), *Handbook of Second Language Assessment* (pp. 241-256). De Gruyter Mouton.

IELTS. (2024). Purpose of the IELTS exam. *IELTS Official Website*. https://www.ielts.org/

IELTS. (n.d.). What can IELTS do for you? IELTS. https://ielts.org/take-a-test/why-choose-ielts/what-can-ielts-do-for-you

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, 38(12), 1217-1218.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 5-55.

Kim, M. K., & Lopez, A. A. (2022). Developing a technology-based classroom assessment of academic reading skills for English language learners and teachers: Validity evidence for formative use. *Languages*, *7*(2), 71. https://doi.org/10.3390/languages7020071

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*, (pp. 1–13). Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds.), *European language testing in a global context,* (pp. 27–48). Cambridge University Press.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183–189. https://doi.org/10.1177/0265532209349468

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment: Challenges and opportunities in opening the black box. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000790

McKim, C. (2023). Meaningful member-checking: A structured approach to member-checking. *American Journal of Qualitative Research, 7*(2), 41-52. https://doi.org/10.29333/ajqr/12973

McNamara, T. (2005). 21st century Shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351-370.

McNamara, T. & Ryan, K. (2011). Fairness versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*, 161-78.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Moore, T., Morton, J., & Price, S. (2015). Construct validity in the IELTS Academic Reading and Writing Tests: A comparison of reading and writing requirements in IELTS test items and in university study. *IELTS Research Reports, 3*, 1-26.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method

implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 533-544.

Pishghadam, R., & Tabataba'ian, M. S. (2011). IQ and test format: A study into test fairness. *Iranian Journal of Language Testing, 1*(1), 17–29.

Rudner, L. M., & Schafer, W. D. (2002). *What teachers need to know about assessment*. National Education Association

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. *Journal of Educational Computing Research, 49*(2), 151-179. https://doi.org/10.2190/EC.49.2.a

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391. https://doi.org/10.1177/026553220101800404

Shohamy, E. (2001b). *The power of tests: A critical perspective on the uses of language tests*. Longman.

Slomp, D., Corrigan, J., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments. *Research in the Teaching of English, 48*(3), 276–302.

Spolsky, B. (1995). *Measuring language: Language testing and assessment*. Cambridge University Press.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education., 12*(3), 275–287. https://doi.org/10.1080/09695940500337249

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2,* 53-55.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21-36.

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

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement, 72*(4), 533-546.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *67*(2), 219–235. https://doi.org/10.1177/0013164407305592

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Weir, C., O'Sullivan, B., Yan, J., & Bax, S. (2007). Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS writing component: Effects and impact. *IELTS Research Report*. https://www.ielts.org/researchers/research-reports

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://doi.org/10.1177/0265532209349465à

Zaferanieh, E. (2023). *Fairness in high-stakes second language tests: A systematic review*. Unpublished manuscript.

Appendices

Appendix A
Questionnaire

The questionnaire's goal is to get feedback from teachers, examiners, and students regarding how fair they think the IELTS writing assessment criteria and procedures are.

Demographic Information:

1. **Role:**
    - Instructor
    - Examiner
    - Learner
    - Both
2. **Years of IELTS experience (if relevant):**
    - 1 year or less
    - 2 years
    - 3 years
    - 4 years
    - 5 years
    - 6 years
    - 7 years or more
3. **What is your native language?** (Open-ended response)
4. **Which country are you currently residing in?** (Open-ended response)

Section A: Perceptions of Fairness in IELTS Writing Assessment (Educators and Examiners)

5. **How clear do you find the IELTS writing assessment criteria?**
    - 1 = Very Unclear
    - 2 = Somewhat Unclear
    - 3 = Neutral
    - 4 = Somewhat Clear
    - 5 = Very Clear
6. **How fair do you think the IELTS writing assessment is for test-takers from diverse cultural and linguistic backgrounds?**
    - 1 = Very Unfair
    - 2 = Somewhat Unfair
    - 3 = Neutral
    - 4 = Somewhat Fair
    - 5 = Very Fair
7. **To what extent do you believe the IELTS writing assessment addresses cultural differences?**
    - 1 = Not at all

- 2 = Slightly
- 3 = Moderately
- 4 = Mostly
- 5 = Completely

8. **How often do you think bias occurs in IELTS writing assessment due to cultural or linguistic differences?**
   - 1 = Never
   - 2 = Rarely
   - 3 = Sometimes
   - 4 = Very Often
9. **In your opinion, do IELTS writing prompts reflect a Western-centric perspective?**
   - 1 = Strongly Disagree
   - 2 = Disagree
   - 3 = Neutral
   - 4 = Agree
   - 5 = Strongly Agree
10. **Do you believe that the emphasis on IELTS writing scores in university admissions, job placements, or immigration decisions is fair?**
    - 1 = Strongly Disagree
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly Agree

Section B: Challenges Faced by Test Takers (Test Takers Only)

11. **How difficult do you find the IELTS writing assessment tasks compared to real-world writing requirements?**
    - 1 = Much More Difficult
    - 2 = Somewhat More Difficult
    - 3 = About the Same
    - 4 = Somewhat Easier
    - 5 = Much Easier
12. **How clearly do you understand the scoring criteria used in the IELTS writing assessment?**
    - 1 = Not at all
    - 2 = Slightly
    - 3 = Moderately
    - 4 = Very Well
    - 5 = Completely
13. **Do you believe the feedback provided after IELTS writing assessments is helpful for improving your writing skills?**
    - 1 = Not at all Helpful
    - 2 = Slightly Helpful
    - 3 = Moderately Helpful
    - 4 = Very Helpful
    - 5 = Extremely Helpful

14. **How do you feel the time constraints in the IELTS writing assessment affect your performance?**
    - 1 = Significantly Hinder My Performance
    - 2 = Somewhat Hinder My Performance
    - 3 = No Impact on My Performance
    - 4 = Somewhat Enhance My Performance
    - 5 = Significantly Enhance My Performance
15. **Do you feel that the cost of the IELTS test limits your opportunities to retake the exam?**
    - 1 = Strongly Disagree
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly Agree
16. **How fair do you find it that your IELTS writing score could impact university admissions, job placements, or immigration opportunities?**
    - 1 = Very Unfair
    - 2 = Somewhat Unfair
    - 3 = Neutral
    - 4 = Somewhat Fair
    - 5 = Very Fair

Section C: Impact of Technology on Fairness in IELTS Writing Assessment

17. **How do you feel the use of a computer-based test (e.g., typing your essay) affects fairness in the IELTS writing assessment?**
    - 1 = Significantly Decreases Fairness
    - 2 = Somewhat Decreases Fairness
    - 3 = No Impact on Fairness
    - 4 = Somewhat Increases Fairness
    - 5 = Significantly Increases Fairness
18. **Do you think typing your essay instead of handwriting it affects your performance in the IELTS writing task?**
    - 1 = Strongly Disagree
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly Agree
19. **How accessible do you find online resources and practice tools for IELTS writing preparation?**
    - 1 = Very Inaccessible
    - 2 = Somewhat Inaccessible
    - 3 = Neutral
    - 4 = Somewhat Accessible
    - 5 = Very Accessible
20. **Do you believe that using technology (e.g., online test preparation, automated feedback) has improved your IELTS writing performance?**

- 1 = Strongly Disagree
- 2 = Disagree
- 3 = Neutral
- 4 = Agree
- 5 = Strongly Agree

21. **Please provide any further feedback or personal experiences regarding the use of technology, fairness, bias, or challenges in the IELTS writing assessment.**
    - (Open-ended response)

Appendix B

Interview Guide for Semi-Structured Interviews

Questions for interviews:

What has been your overall experience with the IELTS test?

What is your opinion of IELTS writing test's overall fairness?

Do you think there is anything about this test that might be made fairer?

Could you elaborate on any particular obstacles or problems you have encountered or you have seen others encounter with the IELTS writing test?

How do these challenges affect perceptions of fairness among test-takers?

Impact of Test Design and Administration:

In what ways do you think the design and administration of the IELTS writing test impact its fairness?

Are there particular policies or practices that you believe contribute to or detract from the fairness of the test?

To what extent in your opinion, does the IELTS writing exam cater to test-takers with different language and cultural backgrounds?

Would you suggest any particular adjustments to better handle these differences?

What part does technology play in ensuring that IELTS writing exam is scored fairly, in your opinion?

Do you think there are any technical resources or techniques that could help make evaluation process more equitable?

What adjustments or modifications would you recommend to improve fairness of IELTS writing assessment based on your experience?

Appendix C
Consent Form

INFORMATION AND CONSENT TO PARTICIPATE IN A RESEARCH STUDY

Study Title:

Exploring Fairness in Second Language Writing Assessments.

Researcher:

- *Elaheh Zaferanieh*
- *Master's student in Educational Technology*

Researcher's Contact Information:

- *5145629756*
- e_zafera@live.concordia.ca or e.zaferanieh@yahoo.com

Faculty Supervisor:

Dr. Julie Corrigan, Department of Education

Faculty Supervisor's Contact Information:

Julie.Corrigan@Concordia.ca

Source of funding for the study:

You are being invited to participate in the research study mentioned above. This form provides information about what participating would mean. Please read it carefully before deciding if you want to participate or not. If there is anything you do not understand, or if you want more information, please ask the researcher.

A. PURPOSE

The purpose of this study is to investigate how people experience and perceive the fairness of the IELTS exam.

B. PROCEDURES

If you participate, you will be asked to fill out the questionnaire and to take part in a semi-structured interview where you may talk about your experiences taking language tests. You may arrange the interview whenever is most convenient for you. It will be done over Zoom and it will last about fifteen minutes. The interviews will be audio-recorded. The questionnaires will be sent in word format via email, and they should be returned by email too, in two or three days. It will last about 15 minutes too.

C. RISKS AND BENEFITS

There are no known risks related to this study. You are free to skip questions or end the interview at any moment, if you are uncomfortable.

Even though there might not be any immediate benefits for you, your involvement will be helpful in expanding our understanding about how a variety of people (test takers, instructors, examiners) experience and perceive the IELTS exam.

D. CONFIDENTIALITY

We will gather the following information as part of this research: your responses to a questionnaire and your experiences mentioned in the interview about the writing section of the IELTS exam. Your answers and your identity will be kept confidential, and we will remove any information that could be used to identify you in any reports stemming from the research.

We will not allow anyone to access the information, except people directly involved in conducting the research. We will only use the information for the purposes of the research described in this form.

Via email, you will receive a summary of your interview transcript to review for accuracy. You can suggest corrections or clarifications to ensure that your responses are accurately represented. After data analysis, useful data will be archived and de-identified. This data may be used for secondary analyses to further explore related research questions or for educational purposes. Any remaining data will be permanently deleted within five years.

We intend to publish the results of the research. However, it will not be possible to identify you in the published results.

E. CONDITIONS OF PARTICIPATION

You do not have to participate in this research. It is purely your decision. If you do participate, you can stop at any time. If you decide to withdraw, you can contact the student researcher or the supervisor within two weeks after providing the consent form, which is before the study's final analysis.

If you withdraw, you can also ask that the information you provided not be used, and your choice will be respected. The information collected from you will not be used, and any digital and paper files containing your information will be securely deleted.

There are no negative consequences for not participating, stopping in the middle, or asking us not to use your information.

F. PARTICIPANT'S DECLARATION

I have read and understood this form. I have had the chance to ask questions and any questions have been answered. I agree to participate in this research under the conditions described.

NAME (please print) _____

SIGNATURE _____

DATE _____

If you have questions about the scientific or scholarly aspects of this research, please contact the researcher. Their contact information is on page 1. You may also contact their faculty supervisor.

If you have concerns about ethical issues in this research, please contact the Manager, Research Ethics, Concordia University, 514.848.2424 ex. 7481 or oor.ethics@concordia.ca

Appendix D

Recruitment letter for IELTS writing assessment study

Invitation to participate in a study on IELTS writing assessment practices

Dear educator/examiner/test-taker,

At Concordia University, we are carrying out a research project named "Exploring Fairness in Second Language Writing Assessments". This study aims to examine the perceptions and experiences of students, instructors, and examiners with writing section of the IELTS exam.

We welcome you to take part if you have any experience with the IELTS writing examination, either as a test-taker, examiner, or instructor. Your observations will help improve our knowledge of what defines fairness in language evaluations.

Participation will involve a questionnaire (15 minutes) and an interview over Zoom (approximately 15 minutes). Your answers will be confidential, meaning that while the researcher will know your identity, your identity will not be evident in any research reports.

In case you have any questions or would need further details, kindly reach out to us at e_zafera@live.concordia.ca or e.zaferanieh@yahoo.com

Please feel free to forward this information to anybody you know who might be interested in taking part. We appreciate your consideration.