English Language Teaching; Vol. 13, No. 1; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

A Comparability Study of Text Difficulty and Task Characteristics of Parallel Academic IELTS Reading Tests

Linyu Liao[1]

[1] Faculty of Arts and Humanities, University of Macau, Macau

Correspondence: Linyu Liao, Faculty of Arts and Humanities, University of Macau, Macau.

Received: November 9, 2019 Accepted: November 29, 2019 Online Published: December 4, 2019

doi: 10.5539/elt.v13n1p31 URL: https://doi.org/10.5539/elt.v13n1p31

Abstract

As a high-stakes standardized test, IELTS is expected to have comparable forms of test papers so that test takers from different test administration on different dates receive comparable test scores. Therefore, this study examined the text difficulty and task characteristics of four parallel academic IELTS reading tests to reveal to what extent the four tests were comparable in terms of text difficulty, construct coverage, response format, item scope, and task scope. The Coh-Metrix-TEA software was used for the text difficulty analyses and expert judgments were used for task characteristics analyses. The results show that the four reading tests were partly comparable in text difficulty, comparable in terms of construct coverage and item scope, but not comparable in terms of response format and task scope. Based on the findings, implications were discussed on test development and future research.

Keywords: comparability, IELTS Reading, text difficulty, task characteristics, parallel tests

1. Introduction

Comparability of parallel forms of standardized tests that are administered on different test dates and test sites is an essential quality that ensures the equivalence of test scores for similar abilities of test takers. This in turn supports the claim of reliability of the test scores and their interpretation. Thus, it is critical to examine the implicit claim that a test has parallel forms that are equivalent in terms of construct coverage, content, text difficulty, and response format for the test's overall fairness.

This study examined four parallel forms of the IELTS (International English Language Testing System developed and administered by Cambridge Language Assessment, the British Council, and the IDP, Australia) academic reading test. As an influential standardised high-stakes test, the IELTS claims explicitly or implicitly that it is consistent across different forms of test papers in terms of text difficulty, construct coverage, response format, item scope, and task scope so that scores obtained from different test administrations are comparable. Such consistency across different forms of a test cannot be taken for granted and needs to be established through well designed and carefully conducted empirical study. Additionally, a review of relevant literature showed that the comparability of parallel forms of tests is under-researched compared to other types of comparability studies.

2. Review of Literature

2.1 Parallel Tests

In this study, the term parallel tests or parallel forms refers to interchangeable versions of a test in terms of construct and content and equivalence in test performance of test takers with similar ability across test administrations. Parallel tests are widely used in large-scale standardized tests such as IELTS, TOEFL, and GRE. As stated earlier, it is vitally important for parallel tests to be comparable or equivalent because incomparability of parallel tests is likely to cause fairness issues and thus erode the value of tests. As Wendler and Walker (2015) pointed out parallel tests must be equivalent to ensure the interchangeability of test scores. However, there is a lack of evidence for the comparability and equivalence of parallel tests and testing agencies are often criticized for not providing such evidence (Bachman et al., 1995; Chalhoub-Deville & Turner, 2000; Spolsky, 1995; Weir & Wu, 2006).

Evidence for the comparability of parallel tests typically include both test construct and content comparability and test score equivalence (Bae & Lee, 2010; Wendler & Walker, 2015). Test score equivalence can be examined by examining test takers' performance on different parallel tests (see Bae & Lee, 2010; Weir & Wu, 2006).

Important as test score equivalence is, it should be noted that such equivalence is about test scores only, but not about tests per se. In order to achieve test score equivalence, test content equivalence should be guaranteed from the outset. Test content equivalence can be measured from different aspects of task characteristics such as text difficulty, construct coverage, and response format (see Bachman, Davidson, & Milanovic, 1996; Kunnan & Carr, 2017).

2.2 Previous Comparability Studies

A few comparability studies are briefly reviewed to understand the scope of this type of research and to draw on good research design and methods. Comparability studies can be roughly classified into four types according to their research foci: (1) comparability between different tests with the same purpose; (2) comparability between different versions of the same test, for example, paper-based and computer-based versions; (3) comparability between parallel forms of the same test; (4) comparability of task types of the same test.

2.2.1 Comparability Between Different Tests

Bachman et al.'s (1995) Cambridge-TOEFL comparability study conducted in the late 1980s was one of the earliest well-known comparability studies. It examined the comparability between FCE (First Certificate English) by Cambridge ESOL (previously known as University of Cambridge Local Examinations Syndicate) and TOEFL (Test of English as a Foreign Language) by ETS (Educational Testing Service, Princeton). The research methods used in this study included a quantitative analysis of test performance data and a qualitative content analysis of test prompts, tasks, and items by expert judges. The study found that the two tests basically measured similar language abilities and there were more similarities than differences between the two tests. This carefully designed research study was later followed partly or fully in other comparability studies.

Kunnan and Carr (2017) recently conducted a comparability study of the reading section of GEPT-A (General English Proficiency Test-Advanced) and iBTOEFL (internet-Based Test of English as a Foreign Language). This study drew on Bachman et al.'s (1995) research design and performed content analysis on reading texts and task analysis in the reading test items, and test performance analysis of the scores on the two tests. The results of the text analysis showed the reading texts on the two tests were comparable in many ways but differed in several key regards.

2.2.2 Comparability Between Different Types of Administrations of the Same Test

Studies on the comparability between paper-based and computer-based versions of the same test have become increasingly popular due to the proliferation of computer-delivered tests. Bridgeman and Cooper (1998) conducted an early large-scale comparability study of this type. They compared the scores of word-processed and handwritten GMAT (Graduate Management Admission test) essays by 3,470 test takers. Results showed that the hand-written essays tended to get higher scores than the word-processed essays.

In the last two decades similar studies have been providing different results. For example, Sawaki (2001) evaluated the construct comparability between traditional paper-based and computer-delivered L2 reading tests. Her review covered a wide range of relevant issues, such as cognitive ability, ergonomics, etc. The study, however, did not draw a clear-cut conclusion about the equivalence of the two types of tests. Choi, Kim, and Boo (2003) also compared paper-based and computer-based versions of TEPS (Test of English Proficiency developed by Seoul National University). Corpus-based content analysis and statistical analyses supported the comparability of the two versions of the test. Other empirical studies have found that test takers tend to perform better on paper-based tests than on computer-delivered tests (e.g., Choi & Tinkler, 2002; Murphy et al., 2003; O'Malley et al., 2005). Detailed discussions on this respect can also be seen in Paek (2005), who noticed divergent findings of comparability studies since 1993. On the one hand, computers were often used to administer traditional tests without significant influence on test scores. On the other hand, many studies found better performance on paper-based tests than on computer-based tests. Paek (2005) believed that as test takers got more familiar with computer interfaces and computer-delivery became more user-friendly, the differences between computer-based and paper-based tests would disappear.

Relevant studies were also conducted by O'Loughlin (1997, 2001), who examined the equivalence of direct and semi-direct (live and tape-based) versions of a speaking test. Overall, the author suggested that the speaking tests in these two methods cannot be substituted for each other because they measured different constructs.

2.2.3 Comparability Between Parallel Forms of the Same Test

Compared with the first two types of comparability studies, the comparability between parallel forms of the same test seems to attract much less attention. One relevant study was conducted by Bachman, Davidson, and Milanovic (1996). They examined the equivalence of six forms of the FEC (First Certificate of English) in terms

of communicative language abilities under test, linguistic characteristics, and response format with newly developed CLA (Communicative Language Ability) using the Test Method Facet framework. The results showed that not all facets were comparable between the parallel tests. This study indicated that parallel forms of tests were not always comparable, which underscores the need for the current study.

However, a more recent study by Weir and Wu (2006) showed that the three parallel forms of GEPTS-I (General English Proficiency Test Speaking-Intermediate) were generally comparable and Forms 2 and 3 could even be considered equivalent at task level. The different results in the above-mentioned two studies show that the comparability of parallel tests varies from case to case. Therefore, comparability studies between parallel forms of the same test should be conducted on an individual basis. To date, there have not been studies on the comparability between parallel forms of IELTS; the current study meets this research gap.

2.2.4 Comparability of Task Types in the Same Test

This type of study too has not been popular. One of the few examples was a study by Hancock (1994), who examined the comparability of multiple-choice questions and constructed-response items of a reading test. Correlation analysis and factor analysis showed that the two types of tasks measured similar constructs. More recently, some studies have been conducted to compare independent and integrated writing tasks in TOEFL iBT. Through think-aloud verbal report and interviews, Plakans (2010) found that the two types of writing tasks were interpreted as similar among some participants and also interpreted as different among others. In addition, Cumming et al. (2005) found significant difference in test takers written discourse in the two types of tasks.

2.3 Current Study and Research Questions

The current study falls into the third type of comparability research, which is under-researched. This type of research is critical as it is important to examine a test agency's claim that their test scores are reliable and, therefore, fair to all test takers. From the perspective of research design, existing comparability studies have mainly used three methods: construct and content analysis of reading texts with corpus-based analytical tools, task analysis based on expert judgments, and statistical analysis of test performance data. This study used the first two methods; the third method, statistical analysis of test scores, was not conducted due to the lack of relevant data. The two specific research questions were:

Research Question 1: To what extent is the content of the four parallel forms of academic IELTS reading comparable in terms of text difficulty?

Research Question 2: To what extent is the content of the four parallel forms of academic IELTS reading comparable in terms of task characteristics such as construct coverage, response format, item scope, and task scope?

3. Research Methods

3.1 Materials

Four IELTS reading tests were used in this study. They were taken from Cambridge English IELTS 12 Academic (Cambridge English, 2017) officially published by Cambridge University Press (CUP). According to CUP, the book includes authentic examination papers of IELTS. Each reading test contained three reading texts and 40 test items. Each reading text included two to three tasks, which altogether consist of 13 or 14 test items. The titles of the reading texts are provided in Appendix A.

3.2 Text Difficulty Analysis

Text difficulty in reading has traditionally been attributed to length of sentences and the size or frequency of words in a text. But today there is growing awareness that there are other linguistic features that contribute to text difficulty such as narrativity and cohesion. The Coh-Metrix Text Easability Assessor (Coh-Metrix-TEA) is an automated tool that quantitatively analyses texts for linguistic and discourse features by examining narrativity, syntax simplicity, word concreteness, referential cohesion, and deep cohesion (Graesser, McNamara, Cai, Conley, & Pennebaker, 2014; McNamara, Graesser, McCarthy, & Cai, 2014). In addition, Coh-Metrix-TEA provides a Flesch-Kincaid grade level (FKGL) readability score based mainly on the length of sentences and the size of words in a text. High indices on the first five measures indicates greater easiness whereas a high readability score on the FKGL indicates greater difficulty.

Figure 1 shows an example of Coh-Metrix-TEA output from an analysis of a reading text. In this example, the FKGL index is 10.4, indicates the text can be understood by 10th grade L1 English students in a U.S. school system. This index is supposed to help teachers, parents, librarians, and others to judge the readability level of various books and texts. This sample text analysis shows percentiles on indicators such as narrativity, syntax

simplicity, and referential cohesion; this indicates that these indicators are increasing text difficulty. On the other hand, high percentiles on two indicators, word concreteness and deep cohesion, shows that these indicators are reducing text difficulty.

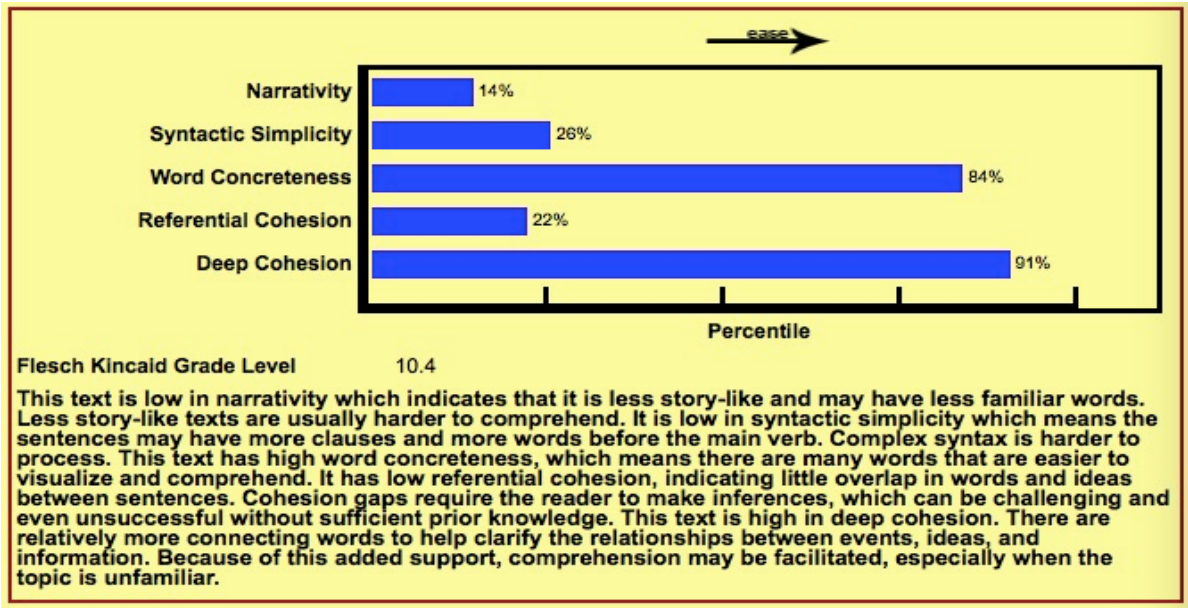

Figure 1. An example of Coh-Metrix-TEA output

3.3 Task Characteristics Analysis

3.3.1 Expert Judgement

The expert judgment method was used for task analysis. Rater 1 was one of the authors of this study, a doctoral student in English Linguistics, with about two years' IELTS related teaching experience. Rater 2 was an English language teacher who had about two years' IELTS related teaching experience and held an MA degree in English Language and Applied Linguistics.

Rater 1 first summarised a preliminary analytical framework for task analysis according to the information on the IELTS website as well as the framework used in the Kunnan and Carr (2017) comparability study. The two raters conducted preliminary analyses; then, they adjusted and finalized the analytical framework. Rater 1 and 2 analysed Reading 1 texts (see Appendix B). As the analytical framework was quite straightforward, a high inter-rater agreement (95%) was achieved. Disagreements regarding ratings were settled through discussion. As the 95% agreement on Reading test 1 (25% of the total data) was much higher than Smagorinsky's (2008) suggestion of a minimum agreement level of 80 over 15% of the data, Rater 1 continued to analyse the remaining three reading tests.

3.3.2 Analytical Frameworks

Task characteristics analyses covered four aspects: construct coverage, response format, item scope, and task scope. The IELTS framework Academic "Reading in detail" for the preliminary analysis of construct coverage was adapted for this study (see Appendix C). After a preliminary analysis of construct coverage, the expert raters found it difficult to distinguish between "locating specific information," "identifying specific information" and "understanding specific information." The researchers felt that test items related to details involved all three abilities. Thus, it was decided to simplify the framework for the analysis of construct coverage by using an umbrella construct "reading for details." The revised framework shown in Table 1 was used for the study.

Table 1. Revised framework for the analysis of construct coverage

Construct component	Task Type*
Reading for main ideas	Matching headings
	Multiple choice
Reading for details	Matching information
	Matching features
	Sentence completion
	Matching sentence endings
	Summary completion
	Notes completion
	Table completion
	Flow-chart completion
	Diagram label completion
	Short-answer questions
	True/False/Not Given
	Multiple choice
Identifying writer's opinions or ideas	Yes/No/Not Given
	Multiple choice

* Sample tasks of each task type are available on the official website of IELTS.

A taxonomy for response format that included matching, blank-filling, multiple-choice, and other task types was also developed. This taxonomy is presented in Table 2.

Table 2. Taxonomy of response format

Response Format	Task Type
Matching	Matching headings
	Matching information
	Matching features
	Matching sentence endings
Blank-filling	Sentence completion
	summary completion
	Notes completion
	Table completion
	Flow-chart completion
	Diagram label completion
Judging	True/False/Not Given
	Yes/No/Not Given
Selecting from multiple choice	Multiple choice with one correct answer or more than one correct answer
Answering short-answer questions	Short-answer questions

The analytical framework for item scope analysis was adapted from Kunnan and Carr (2017). Item scope was rated on a scale from 'very narrow' to 'very broad' according to the amount of key information that gave hints to

getting the correct answer. Item scope was seen as 'very narrow' if the key information was found within a single sentence, 'narrow' if the key information was found within two consecutive sentences in a paragraph, 'moderate' if the key information was found within three consecutive sentences in a paragraph, 'broad' if the key information was found in more than three consecutive sentences in a paragraph or within two paragraphs, and 'very broad' if the key information was found in more than two paragraphs.

Similarly, an analytical framework was developed for task scope analysis according to the number of consecutive paragraphs that include all the answers of the task. Task scope was seen as 'very narrow' if all the answers to the task were found within a single paragraph, 'narrow' if all the answers to the task were found within two consecutive paragraphs, 'moderate' if all the answers of the task were found within three consecutive paragraphs, 'broad' if all the answers of the task were found within four consecutive paragraphs, and 'very broad' if all the answers of the task were found in more than four consecutive paragraphs. Table 3 shows the analytical framework for item scope and task scope analyses.

Table 3. Analytical frameworks for the analyses of item scope and task scope

Scope	Description of Item Scope	Description of Task Scope
Very narrow	Within 1 sentence	Within 1 paragraph
Narrow	Within 2 consecutive sentences in a paragraph	Within 2 consecutive paragraphs
Moderate	Within 3 consecutive sentences in a paragraph	Within 3 consecutive paragraphs
Broad	More than 3 consecutive sentences in a paragraph or within 2 paragraphs	Within 4 consecutive paragraphs
Very broad	More than 2 paragraphs	More than 4 consecutive paragraphs

As Table 3 shows, item scope shows the number of sentences or paragraphs that test takers need to read in order to get the answer of an item, and task scope shows the number of paragraphs that that test takers need to read in order to get the answers of a task. In IELTS reading, each task contains several items. It is possible that both item scope and task scope are narrow or that item scope are narrow but task scope is broad. These cases obviously pose different degrees of difficulty for test takers. Therefore, both item scope and task scope were examined.

Based on the expert judgment, descriptive analysis was used to sort and compare the data. In addition, the chi-square test was conducted to check if there was a statistically significant difference in test item characteristics among the parallel tests where appropriate (i.e., items with no frequency of zero and above 80% of frequencies greater than five).

4. Results

4.1 Text Difficulty Analysis

Table 4 presents descriptive statistics of content analysis based on Coh-Metrix-TEA. According to the results, different texts have different sources of difficulty and the overall text difficulty levels vary across tests. In terms of FKGL, the overall difficulty levels of texts in Test 3 and Test 4 were very similar, ranging from approximately 11 to 15, and the overall text difficulty levels of the texts in the two tests were almost the same (12.5 and 12.4 respectively). In contrast, Test 1 and Test 2 varied greatly in text difficulty. All three texts in Test 2 had a higher difficulty level than those in Test 1, and there was a gap of 3.3 between the two tests in average text difficulty (10.5 versus 13.8). The indices that contribute to the overall text difficulty (i.e., syntax simplicity, word concreteness, referential cohesion and deep cohesion) varied greatly across the texts and across the tests. There was not a clear pattern that could be summarised for these texts on the five indices.

In conclusion, the four tests are only partly comparable in terms of overall text difficulty. Specifically, Test 3 and Test 4 are comparable to each other while Test 1 and Test 2 are not comparable to any other tests in overall text difficulty. However, due to the small sample size, no inferential analysis could be conducted to check if the difference or partial similarity in text difficulty among the tests was statistically significant or not.

Table 4. Descriptive statistics of content analysis

Passage	FKGL	Narrativity	Syntax Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion
T1R1	10.4	14	26	84	22	91
T1R2	11.6	41	45	43	11	97
T1R3	9.4	41	58	13	12	83
T1 Mean	10.5	32	43	46.7	15	90.3
T2R1	16.1	8	52	64	4	66
T2R2	12.6	45	10	64	31	72
T2R3	12.6	25	44	50	19	69
T2 Mean	13.8	26	35.3	59.3	18	69
T3R1	11.5	8	52	94	4	39
T3R2	15.1	23	12	56	60	38
T3R3	10.9	44	40	56	15	76
T3 Mean	12.5	25	34.7	68.7	26.3	51
T4R1	11.6	18	37	90	46	96
T4R2	10.8	26	44	37	4	36
T4R3	14.8	12	18	45	13	74
T4 Mean	12.4	18.7	33	57.3	21	68.7

Note. T1R1: Test 1 Reading 1, T1R2: Test 1 Reading 2 … T4R3: Test 4 Reading 3

FKGL: Flesch Kincaid Grade Level

4.2 Task Characteristics Analysis

4.2.1 Construct Coverage

Table 5 shows the construct coverage of the four reading tests. All of the four tests had the highest percent of items assessing the ability to read for details (62.6%-75%). Tests 1 to 3 had the second highest percent of items assessing the ability to read for main ideas (15%-17.5%) and the least items assessing the ability to identify the writer's opinions or ideas (10%-12.5%). Test 4 was the only exception, with more items assessing the ability to identify writer's opinions or ideas. Another striking feature was that Test 2 and Test 3 had identical construct coverage as they had the same percentage of items for each construct component (reading for main ideas: 17.5%, reading for details: 70%, identifying writer's opinions or ideas: 12.5%). Moreover, the $\chi 2$ test shows that the difference in construct coverage among different tests was not statistically significant ($\chi 2 = 2.21$, $p = 0.90$). Thus, based on these results, it can be concluded that the four IELTS academic reading tests can be seen as comparable in terms of construct coverage.

Table 5. Construct coverage of the four reading tests

Construct Component	Items in Test 1		Items in Test 2		Items in Test 3		Items in Test 4	
	N	%	N	%	N	%	N	%
Reading for main ideas	6	15.0	7	17.5	7	17.5	7	17.5
Reading for details	30	75.0	28	70.0	28	70.0	25	62.5
Identifying writer's opinions or ideas	4	10.0	5	12.5	5	12.5	8	20.0

Note. N: Number of Items; %: Percent of Items

4.2.2 Response Format

Table 6 summarises the information regarding the response format in the four reading tests. The results show that the four tests varied greatly in terms of response format. Test 1 had a clear focus on blank-filling and judging questions, which represented 47.7% and 37.5% respectively. Test 2 focused on matching, with 50% of the items using this format. Test 3 used the most matching and blank-filling items, which altogether accounted for 87.5% of the total test items. Similar to Test 1, Test 4 also emphasised blank-filling and judging questions, but Test 4 had a smaller percent of these two types of questions, each representing 35%. In addition to the difference in the main response format, Test 1 did not have any multiple-choice questions and test 3 did not include judging questions.

Further, judging and multiple-choice items were combined as a group for chi-square test as both could involve guessing. The $\chi 2$ test result shows that the difference in response format among different tests was statistically significant ($\chi 2 = 26.06$, $p < 0.05$). Therefore, it can be concluded that the four IELTS academic reading tests were not comparable in terms of response format.

Table 6. Response format of the four reading tests

Response Format	Items in Test 1		Items in Test 2		Items in Test 3		Items in Test 4	
	N	%	N	%	N	%	N	%
Matching	6	15.0	20	50.0	17	42.5	7	17.5
Blank-filling	19	47.5	7	17.5	18	45.0	14	35.0
Judging	15	37.5	9	22.5	0	0.0	14	35.0
Multiple choice	0	0.0	4	10.0	5	12.5	5	12.5

Note. Short-answer questions are excluded from the table due to no relevant items in the four tests.

4.2.3 Item Scope

Table 7 presents information about the item scope of the four reading tests. According to the Table, the vast majority of the items (87.5%-90%) in each test had a narrow or very narrow scope, which meant between 87.5%-90% of the answers could be found within two consecutive sentences. In particular, items with very narrow scope represented 62.5%-70% of the total items in each test. That is to say, 62.5%-70% of the answers in each test could be answered by understanding one sentence. The chi-square test was conducted with moderate to very broad scope combined into one category. The result showed the difference in item scope among the four tests was not statistically significant ($\chi 2 = 0.94$, $p = 0.99$). Thus, it can be concluded that the four reading tests were comparable in terms of item scope.

Table 7. Item scope of the four reading tests

Item Scope	Items in Test 1		Items in Test 2		Items in Test 3		Items in Test 4	
	N	%	N	%	N	%	N	%
Very narrow	27	67.5	25	62.5	27	67.5	28	70.0
Narrow	9	22.5	10	25.0	9	22.5	7	17.5
Moderate	1	2.5	5	12.5	2	5.0	4	10.0
Broad	3	7.5	0	0.0	2	5.0	1	2.5
Very broad	0	0.0	0	0.0	0	0.0	0	0.0

4.2.4 Task Scope

Table 8 lists the number and the percentage of tasks with different ranges in scope in each test. As the table shows, 86.7% of the tasks in Test 1 had a narrow or very broad scope, each representing 42.9% respectively. Test 2 and Test 3 focused on tasks with a very broad scope, accounting for 66.7% and 57.1% respectively. In Test 4, tasks with moderate scope had the largest percentage, at 37.5%. These results showed that the four IELTS

academic reading tests were not comparable in terms of task scope. However, due to the low frequencies of tasks (all lower than five) with different scope, the chi-square test could not be conducted to check if the difference in task scope among the four tests were statistically significant or not.

Table 8. Task scope of the four reading tests

Task Scope	Tasks in Test 1		Tasks in Test 2		Tasks in Test 3		Tasks in Test 4	
	N	%	N	%	N	%	N	%
Very narrow	0	0.0	0	0.0	1	14.3	1	12.5
Narrow	3	42.9	2	22.2	1	14.3	1	12.5
Moderate	0	0.0	1	11.1	0	0.0	3	37.5
Broad	1	14.3	0	0.0	1	14.3	1	12.5
Very broad	3	42.9	6	66.7	4	57.1	2	25.0

Note. The total number of tasks in different tests are not the same. Test 1: 7 tasks; Test 2: 9 tasks; Test 3: 7 tasks; Test 4: 8 tasks.

5. Discussion and Conclusion

This study examined the comparability of four IELTS academic reading tests. Two research questions were posed regarding the comparability: text difficulty and task characteristics (in terms of construct coverage, response format, item scope and task scope). Based on the analyses, the research questions could be answered as follows:

Research Question 1: To what extent are the four parallel forms of academic IELTS reading comparable in terms of text difficulty?

The four tests were partly comparable in terms of text difficulty. More specifically, Test 3 and Test 4 were highly comparable to each other while Test 1 and Test 2 were not comparable to any other tests in terms of text difficulty.

Research Question 2: To what extent are the four parallel forms of academic IELTS reading comparable in terms of task characteristics such as construct coverage, response format, item scope, and task scope?

The four parallel forms of IELTS academic reading were comparable in construct coverage and item scope, but not comparable in response format or task scope.

In general, the study found that the four parallel IELTS academic reading tests had different levels of comparability in different aspects. In this study, there was a big gap in reading text difficulty between some tests, which would likely influence test taker performance. The lack of comparability in task scope may also cause unfairness and have a negative impact on test takers. Since a task with a narrower scope requires a smaller amount of reading, tests that have more tasks with a narrower scope may favour their test takers and disadvantage candidates taking tests with broader task scope. The incomparability of the response format too is another factor that may pose a threat to test score reliability and, thus, test fairness. For example, since judging and multiple-choice questions are likely to involve guessing, test takers have chances to get the correct answer without reading the questions or the texts. Studies have shown that response format has a significant influence on test performance (e.g., Bachman et al., 1996; Kobayashi, 2002). Therefore, it is suggested that IELTS should attempt to equalize the number of items in various response formats in the parallel reading tests.

The finding of this study is different from Weir and Wu's (2006) study, in which they found that parallel forms of GEPTS-I were generally comparable. In this study, no such overall comparability in parallel academic IELTS reading tests was found. This finding is in line with Bachman et al.'s (1996) finding that not all facets of parallel tests were comparable. These different results of comparability studies show that some tests have more comparable parallel tests while some others do not. They also indicate that it is possible to improve the comparability of parallel tests. Since the lack of comparability of the parallel tests in certain aspects may cause test score unreliability and, thus, unfairness to test takers, the incomparability of parallel academic IELTS reading tests seems to be quite concerning, especially when IELTS is an international large-scale high-stakes test. Therefore, the aim for test developers would be to attempt to produce more comparable parallel tests and routinely examine the comparability of parallel tests.

Although this research was carefully conducted, there are some limitations to the study. First, the results from this study may lack generalisability due to the small sample size of test materials reviewed; only four parallel academic IELTS reading tests, which could represent only a small portion of IELTS parallel tests. Second, the findings for text difficulty are solely based on Coh-Metrix indices, especially Flesh-Kincaid grade level, which is less predictive than the results generated from natural language processing tools (Crossly et al., 2017). Cross-validation of these findings with other measures of text difficulty or more advanced NLP tools would enhance the value of these findings. In addition, this study only examined text difficulty at linguistic level. We are aware that other factors such as topic familiarity or background knowledge may also influence text difficulty, but these factors were not investigated in this study. Finally, without analysing test performance data, it is not known to what extent and how differences in text difficulty and task characteristics in parallel tests may influence test performance or score comparability. When relevant test performance data are available, further research is suggested to investigate the effect of these individual factors as well as the interaction effect of multiple factors on score comparability of parallel tests. Such analyses would offer a more comprehensive understanding of parallel academic IELTS reading tests.

References

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2*, 1-34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*(2), 125-150. https://doi.org/10.1177/026553229601300201

Bachman, L. F., Davidson, F., Ryan, K., Choi, I-C. (1995). *An investigation of the comparability of the two tests of English as a foreign language: the Cambridge-TOEFL comparability study*. Cambridge, UK: Cambridge University Press.

Bae, J., & Lee, Y. S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing, 28*(2), 155-177. https://doi.org/10.1177/0265532210382446

Bridgeman, B., & Cooper, R. (1998). *Comparability of scores on word-processed and handwritten essays on the graduate management admissions test*. Princeton, NJ: Research report 143: Educational Testing Service.

Cambridge English (2017). *Cambridge English IELTS 12 Academic*. Cambridge, UK: Cambridge University Press.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28*(4), 523-539. https://doi.org/10.1016/S0346-251X(00)00036-1

Choi, I-C., Kim, K. S., Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295-320. https://doi.org/10.1191/0265532203lt258oa

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. In Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43. https://doi.org/10.1016/j.asw.2005.02.001

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5-6), 340-359. https://doi.org/10.1080/0163853X.2017.1296264

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210-229. https://doi.org/10.1086/678293

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education, 62*(2), 143-157. https://doi.org/10.1080/00220973.1994.9943836

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220. https://doi.org/10.1191/0265532202lt227oa

Kunnan, A. J., & Carr, N. T. (2017). A comparability study between the General English Proficiency Test – Advanced and the Internet-Based Test of English as a Foreign Language. *Language Testing in Asia, 7*(1), 7-17. https://doi.org/10.1186/s40468-017-0048-x

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Murphy, P. K., Long, J. F., Holleran, T. A., & Esterly, E. (2003). Persuasion online or on paper: a new take on an old issue. *Learning and Instruction, 13*(5), 511-532. https://doi.org/10.1016/S0959-4752(02)00041-5

O'Loughlin, K. (1997). *The comparability of direct and semi-direct speaking tests: a case study*. Unpublished Ph.D. thesis. Melbourne: University of Melbourne.

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, UK: Cambridge University Press.

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M. C., & Sanford, E. E. (2005). *Comparability of a paper based and computer-based reading test in early elementary grades*. In AERA Division D Graduate Student Seminar, Montreal, Canada.

Paek, P. (2005). *Recent trends in comparability studies*. Pearson Educational Measurement.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185-194. https://doi.org/10.5054/tq.2010.215251

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*, 38–59.

Smagorinsky, P. (2008). The method section as conceptual epicenter in constructing social science research reports. *Written Communication, 25*(3), 389-411. https://doi.org/10.1177/0741088308317815

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, UK: Oxford University Press.

Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing, 23*(2), 167-197. https://doi.org/10.1191/0265532206lt326oa

Wendler, C. L. W., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. Lane, M. R. Raymond, and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 433-449). London: Routledge.

Appendix A

Titles of Reading Texts

Test	Text	Titles
Test 1	Text 1	Cork
	Text 2	Collecting as a hobby
	Text 3	What's the purpose of gaining knowledge
Test 2	Text 4	The risks agriculture faces in developing countries
	Text 5	The lost city
	Text 6	The benefits of being bilingual
Test 3	Text 7	Flying tortoises
	Text 8	The intersection of health sciences and geography
	Text 9	Music and emotions
Test 4	Text 10	The history of glasses
	Text 11	Bring back the big cites
	Text 12	UK companies need more effective boards of directors

Appendix B

A Sample Task Analysis

Item	Task Type	Construct Coverage	Response Format	Item Scope	Task Scope
1	Matching information	Reading for details	Matching	Very narrow	Very broad
2				Narrow	
3				Narrow	
4	Matching features			Narrow	Very broad
5				Very narrow	
6				Very narrow	
7				Very narrow	
8				Very narrow	
9				Very narrow	
10	Multiple choice		Multiple choice	Very narrow	Moderate
11				Very narrow	
12				Moderate	
13				Very narrow	

Appendix C:

IELTS Framework for Preliminary Analysis of Construct Coverage

Construct component	Task Type
Recognising main ideas of paragraphs or sections	Matching headings
	Multiple choice
Locating specific information	Matching information*
	Matching features*
	Sentence completion
	Multiple choice
Understanding specific information	Matching sentence endings
	Summary completion
	Notes completion
	Table completion
	Flow-chart completion
	Diagram label completion
	Short-answer questions
	Multiple choice
Identifying specific information	True/False/Not Given
	Multiple choice
Identifying writer's opinions or ideas	Yes/No/Not Given
	Multiple choice

* Matching information: Finding information in which paragraph or section

* Matching features: Matching findings to researchers, events to historical periods, etc.

Source: https://www.ielts.org/about-the-test/test-format