TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

Text Complexity of Cambridge-delivered IELTS Academic Reading Tests: Comparability with IELTS Academic Reading Practice Tests from Other Publishers

August 2024 – Volume 28, Number 2
https://doi.org/10.55593/ej.28110a4

Huu Thanh Minh Nguyen
University of Foreign Language Studies, The University of Danang
<nhtminh@ufl.udn.vn>

Nguyen Van Anh Le
University of Foreign Language Studies, The University of Danang
<lnvanh@ufl.udn.vn>

Abstract

Comparing language tests and test preparation materials holds important implications for the latter's validity and reliability. However, not enough studies compare such materials across a wide range of indices. Therefore, this study investigated the text complexity of IELTS academic reading tests (IRT) and IELTS reading practice tests (IRPrT). Fine-grained quantitative analyses were undertaken to delineate measures of lexical, syntactic, and discourse complexity across a corpus of 108 IRT and 108 IRPrT published by Pearson, Macmillan, and Cengage Learning. The results suggest little difference between IRT and IRPrT at the lexical level; however, there were significant differences in some measures of syntactic and discourse level complexity. The findings bear implications for stakeholders including learners as test takers, instructors, material developers, and language testing researchers. We interpret this to mean that while IRPrT materials are lexically conducive to the practice for IRT and can provide a similar experience, the IRPrT do show some differences in the amount of subordination and idea repetition at the discourse level. Therefore, instructors and learners may seek to supplement practice with these structures when preparing for IRT , and the designers of such practice materials should consider aligning these factors in the future.

***Keywords*:** Language testing, Reading tests, Reading practice test materials, Lexical text complexity, Syntactic text complexity, Discourse text complexity

Readability research and text-leveling schemes have emphasized the need to delineate text complexity to interpret the interaction between the reader and the text (Mesmer et al., 2012). Text complexity refers to the "text elements that can be analyzed, studied or manipulated" (Mesmer et al., 2012, p. 236). According to Snow's (2002) RAND model of reading comprehension, text elements including vocabulary, syntax, and discourse affect how readers construct different representations of a text, including the surface code (i.e., the exact wording), the text base (i.e., idea units representing textual meaning), and the mental models (i.e., the

way of processing textual information). For this reason, understanding text complexity at lexical, syntactic and discourse levels is necessary for research on L2 reading to better understand the relationship between texts and readers (Alderson, 2000; Guthrie et al., 2013).

High-stakes English language proficiency tests (ELPT) are important to test takers as gate-keeping tools. One such test is the International English Language Testing System (IELTS), which is often used for selection in academic contexts (Pearson, 2019). Because of its importance, preparation for the IELTS has become important, and increasing attention has been given to practice test materials, which are generally considered helpful to test takers (Kirby, 2016; O'Sullivan et al., 2019). This has given rise to an international industry of practice materials for test preparation, but since such materials are not created by the test designers themselves, there are questions about how comparable the IELTS academic reading practice tests (IRPrT) are to the actual IELTS academic reading tests (IRT) (Bachman et al., 1996; Kunnan & Carr, 2017).

However, to our knowledge, no studies have been conducted to examine whether IRPrT have similar text complexity to IRT preparation. This is an important potential gap in the literature because incongruencies between practice test materials and the actual test could potentially negatively affect learners (Green, 2007). Therefore, this study seeks to compare the text complexity of IRT with IRPrT at the lexical, syntactic and discourse levels to uncover any significant differences so that educators can make informed decisions about the use of practice materials and so that the designers of such materials can potentially revise them if necessary.

Literature Review

Text Complexity and L2 Reading Comprehension

L2 readability research has been viewed from two main strands. One strand is situated in traditional readability formulas that generally measure the number of words per sentence (i.e., sentence difficulty index), the number of syllables per words (i.e., word difficulty index), and word frequency (Brown, 1998; Greenfield, 1999). However, traditional readability formulas are criticized for being restrained to measures at the word level, given that L2 readability formulas are sensitive to other elements such as syntax and rhetorical organization (Carrell, 1987). Because of this limitation, the other strand has been motivated by more recent formulas that transcend the word level to demonstrate the relationships between textual elements. According to Crossley et al. (2008), the readability formula is not only constructed of word frequency but comprises such additional measures as syntax across sentences, and cohesion at the discourse level.

Text complexity shows a negative correlation with reading comprehension (Yang et al., 2021), although it is potential to support learner engagement with a text (Fulmer et al., 2015). Several empirical studies have revealed the relationships between lexical, syntactic, and discourse features of texts and reading comprehension. In terms of the lexical level, lexical sophistication, diversity and density are the three most widely examined properties (Read, 2000). Lexical sophistication is often considered a measure of what percentage of words the learner is likely to know (Laufer & Nation, 1995; Nation, 2006). Several works have suggested that readers need to be able to understand somewhere between 95 and 98% of the words or word families in a text in order to understand it (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). Aside from lexical sophistication, higher lexical diversity, i.e., the amount of different words in the text, and density, i.e., the amount of content words, contributes to increasing difficulty in reading comprehension (Read, 2000).

Other studies have shown that the syntactic features of a text also impact its complexity (e.g., Perfetti & Stafura, 2014). For example, Shiotsu and Weir (2017) suggests that readers' syntactic knowledge accounts for more variance in the L2 reading comprehension level than lexical knowledge. This is because limited syntactic knowledge may impede the ability to understand a text despite the ability to understand meaning of single words within it (Tong et al., 2024). In addition to this, more complex sentences and grammatical structures make the sentences more difficult to be parsed, thereby possibly increasing text complexity beyond comprehension (Mesmer et al., 2012; Kyle, 2016).

Cohesion has also been suggested as the most prominent discourse level feature to impact text complexity in L2 readability research. In general, more cohesive texts are easier to be comprehended (Gernsbacher, 2013; Graesser et al., 2004), because complex mental representations and recall are not required (Ehrlich, 1991). The use of cohesive devices is conducive to establishing textual coherence (Goldman & Rakestraw, 2000); therefore, texts that are coherent at both sentence and global levels aid readers' memory of text information, increasing comprehension (Koda, 2005). However, increased text cohesion is generally accompanied by the inclusion of more information (Beck et al., 1991), which is associated with "increased text length, density, and complexity,' requiring readers to 'process larger amounts of text-based information" (Ozuru et al., 2009, p. 229).

Text Comparability between IRT and IRPrT

Examining the comparability between reading tests and practice test materials is significant due to washback, i.e., "the effects of tests on the teaching and learning directed toward them" (Green, 2006, p. 334). According to Green's (2007) model of washback direction, washback is positive if there is a consistency "between test design and skills developed by a curriculum or required in a target language use domain" (Green, 2006, p. 339). However, if there is a large gap between the test design in terms of format, content, and complexity and what is learned during test preparation, negative washback, such as construct-irrelevant variance, may occur (Messick, 1989). Therefore, examining the comparability of texts in authentic and practice tests can provide insights into how accurately they can reflect future test performance, and whether or not they are providing sufficiently realistic reading materials in terms of text complexity.

To our knowledge, there has been no investigation into IRPrT, except for Everett and Colman (2003). They examined the appropriateness of content in the listening and reading components of commercially produced IELTS practice tests dating from 1996 to 1998 and simply found that the vocabulary used in the IRPrT in their study varied from being unfamiliar and difficult to familiar while sentence structures ranged from complicated to uncomplicated. However, Everett and Colman (2003) only examined lexical and syntactic levels and did not look at cohesion or other discourse level factors. Furthermore, their study has become somewhat dated as it was conducted on much older materials and was limited to the technology available at the time, i.e., since such time many advances have been made in natural language processing that allow for more text features to be calculated automatically.

Given the lack of recent research in this area, we believe it is important to reexamine the text complexity of the reading passages of test preparation materials and actual reading test passages to investigate how valid more current materials are when viewed from a wider lense of text complexity. Accordingly, this study addresses the following research questions:

1. To what extent are IRT comparable to IRPrT in terms of lexical text complexity?
2. To what extent are IRT comparable to IRPrT in terms of syntactic text complexity?
3. To what extent are IRT comparable to IRPrT in terms of discourse text complexity?

Methodology

Research Design

This study employed two corpora of IRT and IRPrT. The IRT were obtained from Cambridge IELTS series 9-17 published by Cambridge English Assessment. The IRPrT were extracted from the test preparation materials of three publishers:

(1) Pearson – IELTS Practice test Plus (Jakeman & McDowell, 2001), IELTS Practice test Plus 2 (Terry & Wilson, 2005), IELTS Practice test Plus 3 (Matthews & Salisbury, 2011)
(2) Macmillan – IELTS Test Builder 1 (McCarter & Ash, 2008), IELTS Test Builder 2 (McCarter, 2008)
(3) Cengage Learning – Exam Essential Practice Tests: IELTS 1 (Harrison & Whitehead, 2015), Exam Essential Practice Tests: IELTS 2 (Gough & Hutchison, 2015)

The reading texts in both IRT and IRPrT vary in genres. Each corpus comprises 108 texts, with the average number of tokens in each text and the total number of tokens in total being relatively comparable (Table 1). It can therefore be argued that the compilation of both corpora was balanced and representative (McEnery, 2006).

Table 1. IRT and IRPrT corpora

Corpus	No. texts	Average tokens per passage	Total tokens
IRT	108	872	94,247
IRPrT	108	853	92,230

Data Collection Instruments

Lexical text complexity. Measures of lexical text complexity encompass lexical sophistication, diversity, and density (Michel, 2017). Lexical sophistication is traditionally measured by VocabProfilers to reveal the percentage of words in different frequency bands in a text (Cobb, 2009). Kim et al. (2018) extended the analysis of lexical sophistication beyond word frequency by employing the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015). Through TAALES, Kim et al. (2018) included additional domains of lexical sophistication such as word range, contextual distinctiveness, word neighborhood, academic language, and so forth. Lexical diversity is traditionally measured by the type-token ratio (TTR) as the ratio of the number of different words (types) to the total number of words (tokens) (Read, 2000). As text length may influence the TTR, more robust indices were developed to measure lexical diversity. Later, Kyle et al. (2021) devised the Tool for the Automatic Analysis of Lexical Diversity (TAALED) that incorporates a wide range of measures of lexical diversity, namely the classic TTR, MTLD, MATTR, HD-D, and so forth. Lexical density, i.e., the percentage of content words in a text (Fang & Pace, 2013), is also calculated by TAALED for both types and tokens. Table 2 and 3 detail lexical sophistication, diversity and density measures used in the study and adapted from Yu (2021).

Table 2. Lexical sophistication measures from TAALES

Category	Index Name	Description
Text coverage	3000_level	The percentage of words in the most frequent 3000-word level
	5000_level	The percentage of words in the most frequent 5000-word level
Word frequency "BNC_Written_Freq_"	AW_Log	BNC Written Frequency AW Logarithm
	CW_Log	BNC Written Frequency CW Logarithm
	FW_Log	BNC Written Frequency FW Logarithm
Word range "BNC_Written_Range_"	AW	BNC Written Range AW
	CW	BNC Written Range CW
	FW	BNC Written Range FW
Academic language	All_AWL_Normed	Academic Word List All
Word recognition norms	LD_Mean_RT_Zscore	Lexical Decision Time (z-score)
	LD_Mean_Accuracy	Lexical Decision Accuracy
	WN_Zscore	Word Naming Response Time (z-score)
	WN_Mean_Accuracy	Word Naming Response Accuracy
Contextual distinctiveness	lsa_average_all_cosine	LSA Contextual Distinctiveness (all cosine)
Age of exposure	aoe_inverse_average	LDA Age of Exposure (inverse average)

Table 3. Lexical density and diversity measures from TAALED

Category	Index Name	Description
Lexical density "lexical_density_"	types	Content word types (*N*) divided by word types (*N*)
	tokens	Content word tokens divided by word tokens
TTR "simple_ttr_"	aw	TTR for all word types
	cw	TTR for content words
	fw	TTR for function words
MATTR "mattr50_"	aw	Moving 50-word Average TTR of all words
	cw	Moving 50-word Average TTR of content words
	fw	Moving 50-word Average TTR of function words
MTLD original "mtld_original_"	aw	Average number of all tokens required to reach TTR >= .720
	cw	Average number of all content word tokens to reach TTR >= .720
	fw	Average number of all function word tokens to reach TTR >= .720
MTLD-MA-Wrap "mtld_ma_wrap_"	aw	Moving Average of MTLD (all words)
	cw	Moving Average of MTLD (content words)
	fw	Moving Average of MTLD (function words)

Syntactic text complexity. Common measures of syntactic complexity include the mean length of clause (MLC), T-unit [1] (MLTU), and sentence (MLS) (Vajjala & Meurers, 2013). The number of dependent clauses, complex T-units and elaborated phrasal structures (e.g., complex nominals, verb phrases) per clause, T-unit and sentence has also been employed to measure syntactic complexity (Lu, 2010; Ortega, 2003; Yu, 2021). Lu (2010) created the L2 Syntactic Complexity Analyzer (L2SCA) that incorporates 14 syntactic complexity measures subsumed into five categories: (a) length of production unit, (b) sentence complexity ratio, (c) the amount of subordination, (d) the amount of coordination, and (e) particular syntactic structures. Table 4 details measures of syntactic complexity in this study.

Table 4. Syntactic complexity measures from L2SCA

Category	Measure	Index Name
Length of the production unit	Mean length of clause	MLC
	Mean length of sentence	MLS
	Mean length of T-unit	MLT
Sentence complexity	Sentence complexity ratio	C/S
Subordination	T-unit complexity ratio	C/T
	Complex T-unit ratio	CT/T
	Dependent clause ratio	DC/C
	Dependent clause per T-unit	DC/T
Coordination	Coordinate phrases per clause	CP/C
	Coordinate phrases per T-unit	CP/T
	Sentence coordination ratio	T/S
Particular structures	Complex nominals per clause	CN/C
	Complex nominals per T-unit	CN/T
	Verb phrases per T-unit	VP/T

Kyle (2016) extended the examination of syntactic complexity to include phrasal complexity (i.e., absolute complexity; Bulté & Housen, 2012) to fully unfold the syntactic features of a text, and created the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASC) to automatically calculate 190 indices of fine-grained clausal and phrasal complexity. In this study, we adopted 15 indices that calculate the average number of particular structures per clause. Auxiliary verbs, bare noun phrase temporal modifiers, negation and discourse markers, existential "there", parataxis, modals, agents, passive auxiliaries, passive clausal and nominal subjects, phrasal verb particles, and undefined dependents are all structures removed from analyses because they contribute less to clausal complexity than other structures measured by TAASSC, as justified below:

- Auxiliary verbs are functional elements that only marginally increase clausal complexity because they are features of obligatory inflection which reflect grammatical rather than lexical meanings (Biber et al., 2014).
- Bare noun phrase temporal modifiers often behave similarly to functional elements signifying time rather than demonstrating syntactic patterns.
- Negation markers and discourse markers are better classified as features of textual cohesion rather than clausal complexity (Halliday & Hasan, 2014).
- The existential "there" is a fixed construction that does not demonstrate syntactic flexibility and instead exemplify an ideational metafunction of text rather than interpersonal adaptations (Biber et al., 2013).
- Parataxis is ambiguous, “sometimes competing with coordination, sometimes with subordination, for the same semantic niche in language” (Hoeksema & Napoli, 1993, p. 291), and therefore is rather limited to additive but disjointed structures.
- Modals indicate mood, which is more of a morphological structure that expresses viewpoint than a structure indicating syntactic sophistication (Bardovi-Harlig, 2000).
- Agents and passive auxiliaries are also more morphological than syntactic. Agents demonstrate relational processes rather than clausal transformations according to Halliday's transitivity system (Thompson, 2014). Passive auxiliaries accompany the passive construction rather than exemplifying complexity in their own right.
- Passive clausal and nominal subjects are byproducts of passivization - the passive itself is the sophisticated structure and already represented elsewhere.
- Phrasal verb particles are minimal functional elements that add meaning and therefore are better understood as increasing semantic complexity rather than syntactic complexity (Garniner & Schmitt, 2015; Spring, 2019). Halliday (2004) characterizes particles as instantiating circumstantial features of verb group rather than increasing complexity.
- Undefined dependents are an artifact of parsing errors rather than sophisticated syntax and such parsing inaccuracies do not genuinely reflect a learner's syntactic competence (Briscoe et al., 2010).

We also used 15 indices of phrasal complexity related to: (a) the average number of dependents per each of the seven noun phrase types and (b) the occurrence of particular independent types regardless of the nominal phrases they occur in. Table 5 details measures of clausal and phrasal complexity employed in this study.

Table 5. Clausal and phrasal complexity from TAASSC

Index Name	Description	Example (Kyle, 2016, p. 55)
Clausal complexity		
acomp	Adjective complement	*She looks [beautiful]$_{acomp}$*
advcl	Adverbial clauses	*The accident happened [as night fell]$_{advcl}$*
advmod	Adverbial modifier	*[Accordingly]$_{advmod}$, I ate pizza*
ccomp	Clausal complement	*I am certain [that he did it]$_{ccomp}$*
cc	Clausal coordination	*Jill runs and [Jack jumps]$_{cc}$*
conj	Conjunction	*He runs and [jumps]$_{conj}$*
mark	Subordinating conjunction	*Forces engaged in fighting [after]$_{mark}$ insurgents attacked*
pcomp	Prepositional complement	*They heard about [you missing classes]$_{pcomp}$*
csubj	Clausal subject	*[What she said]$_{csubj}$ is not true*
xsubj	Controlling subject	*[Tom]$_{xsubj}$ likes to eat fish*
nsubj	Nominal subject	*The [baby]$_{nsubj}$ is cute*
dobj	Direct object	*She gave me [a raise]$_{dobj}$*
iobj	Indirect object	*She gave [me]$_{iobj}$ a raise*
ncomp	Nominal compliment	*He is [a teacher]$_{ncomp}$*
xcomp	Open clausal compliment	*I am ready [to leave]$_{xcomp}$*
Phrasal complexity		
nsubj_deps	Dependents per nominal subject	*[[The]$_{deps}$ man [in the red hat]$_{deps}$]$_{nsubj}$ gave the tall man the money.*
ncomp_deps	Dependents per nominal compliment	*He is [[a]$_{deps}$ [tall]$_{deps}$ man]$_{ncomp}$*
dobj_deps	Dependents per direct object	*The man in the red hat gave the tall man [[the]$_{deps}$ money]$_{dobj}$*
iobj_deps	Dependents per indirect object	*The man in the red hat gave [[the]$_{deps}$ [tall]$_{deps}$ man]$_{iobj}$ the money*
pobj_deps	Dependents per prepositional object	*The man in [[the]$_{deps}$ [red]$_{deps}$ hat]$_{pobj}$ gave the tall man the money*
det_nominal	Determiners per nominal phrases	*[The]$_{det}$ man in [the]$_{det}$ red hat gave [the]$_{det}$ tall man [the]$_{det}$ money*
amod_nominal	Adjective modifiers per nominal phrases	*The man in the [red]$_{amod}$ hat gave the [tall]$_{amod}$ man the money*
prep_nominal	Prepositional phrases per nominal phrases	*The man [in the red hat]$_{prep}$ gave the tall man the money*
poss_nominal	Possessives per nominal phrases	*That is [her]$_{poss}$ red car*
vmod_nominal	Verbal modifiers per nominal phrases	*I don't have anything [to say]$_{vmod}$ to you*
nn_nominal	Nouns as modifiers per nominal phrases	*[Oil]$_{nn}$ prices are rising*
rcmod_nominal	Relative clause modifiers per nominal phrases	*I saw the man [you love]$_{rcmod}$*
advmod_nominal	Adverbial modifiers per nominal phrases	*We will drive the red car [tomorrow]$_{advmod}$*
conj_and_nominal	Conjunctions "and" per nominal phrases	*Jack [and]$_{conj\_and}$ Jill*
conj_or_nominal	Conjunctions "or" per nominal phrases	*Jack [or]$_{conj\_or}$ Jill*

Discourse text complexity. Coh-Metrix is commonly used to analyze cohesion among lexical, syntactic, and semantic properties of texts (Graesser et al., 2004). However, Coh-Metrix has a limited number of indices and does not allow for batch processing. Crossley et al. (2016), therefore, introduced the Tool for the Automatic Analysis of Cohesion (TAACO) that enables batch processing and the examination of local (i.e., sentence-level), global (paragraph-level), and overall text (i.e., text-level) cohesion. Table 6 details measures of cohesion in this study.

Table 6. Cohesion measures from TAACO

Index Name	Description
Lexical overlap (Local and global) "adjacent_overlap_"	
2_all_sent	Lemma [2] types that occur at least once in the next two sentences
2_argument_sent	Noun and pronoun lemma types that occur at least once in the next two sentences
binary_all_sent	Sentences with any lemma types overlapping with the next sentences
binary_argument_sent	Sentences with any noun and pronoun lemma overlapping with the next sentences
2_all_para	Lemma types repeated between paragraphs
2_argument_para	Noun and pronoun lemma types repeated between paragraphs
binary_all_para	Paragraphs with any lemma types overlapping with the next paragraphs
binary_argument_para	Paragraphs with any noun and pronoun lemma types overlapping with the next paragraphs
Semantic overlap (Local and global)	
lsa_1_all_sent	Average latent semantic analysis cosine similarity between all adjacent sentences (with a one-sentence interval)
lsa_2_all_sent	"" (with a two-sentence interval)
lsa_1_all_para	"" all adjacent paragraphs (with a one-paragraph interval)
lsa_2_all_para	"" (with a two-paragraph interval)
Connectives (Text cohesion)	
basic_connectives	Basic connectives (e.g., for, and, or)
all_demonstratives	Demonstratives (e.g., this, that, these, those)
all_additive	Additive connectives (e.g., after all)
all_logical	Logical connectives (e.g., consequently)
Givenness [3] (Text cohesion)	
repeated_content_lemma	Content words repeated at least once divided by all words in the text
repeated_content_ and_pronoun_lemma	Content words and third person pronouns repeated at least once divided by all words in the text

Data Analysis

TAALES (version 2.2), TAALED (version 1.4.1), TAASSC (version 1.3.8, including all the indices in L2SCA), and TAACO (version 2.1.3) were employed to generate data on all measures of text complexity at the lexical, syntactic and discourse levels. The datasets were compiled into spreadsheets, enumerating results per text across IRT and IRPrT sources.

Shapiro-Wilk tests were conducted on each measure to check for normality. Independent t-tests were used when the measure was found to be normal in both corpora, and a Mann-Whitney test was used when the measure was not found to be normal in either one or both corpora (Sainani, 2012). Effect size was measured as Cohen's *d* for parametric measures and Spearman's Rho *rs* for non-parametric measures. These were interpreted according to Plonsky and Oswald (2014).

Findings

Lexical Text Complexity

Table 7 shows the results of the comparisons of lexical sophistication between the two corpora. IRT and IRPrT did not significantly differ in text coverage for the 3,000-word and 5,000-word levels. There was also no significant difference in word frequency and word range between

IRT and IRPrT. In terms of academic language, IRT and IRPrt had no significant difference in the frequency of use for academic words and phrases. There was also no significant difference between IRT and IRPrT regarding word recognition, contextual distinctiveness and age of exposure.

Taken together, IRT and IRPrT indicated no significant difference in lexical sophistication, density and diversity. It is therefore concluded that IRT and IRPrT do not differ in terms of text complexity at the lexical level.

Table 7. Comparisons of measures of lexical sophistication

Category	Indices	*M* (*SD*)		Significance Testing
		IRT	IRPrT	
Text coverage	3000_level	110.62 (2.468)	106.38 (2.507)	z = -.499, p = .62, rs = .03
	5000_level	111.23 (1.546)	105.77 (1.701)	z = -.642, p = .52, rs = .04
Word frequency	AW_Log	-.23 (.085)	-.22 (.091)	t = -.953, p = .34, d = -.70
	CW_Log	-1.03 (.120)	-1.01 (.124)	t = -1.524, p = .13, d = -.21
	FW_Log	1.05 (.073)	1.03 (.068)	t = -1.313, p = .19, d = .17
Word range	AW	69.62 (2.714)	70.10 (2.901)	t = -1.234, p = .22, d = .18
	CW	51.72 (4.061)	52.46 (4.154)	t = -1.321, p = .19, d = .17
	FW	110.19 (.544)	106.81 (.497)	z = -.396, p = .69, rs = .03
Academic language	All AWL	110.47 (.028)	106.53 (.025)	z = -.396, p = .69, rs = .03
Word recognition norms	LD_RT	-.533 (.018)	-.532 (.02)	t = -.087, p = .93, d = .05
	LD_Acc.	114.84 (.004)	102.16 (.004)	z = -1.497, p = .14, rs =.10
	WN_Zscore	-.488 (.019)	-.492 (.02)	t = 1.484, p = .14, d = .21
	WN_Acc.	115.01 (.002)	101.99 (.002)	z = -1.560, p = .12, rs = .11

Table 8 shows the results of the comparisons of lexical density between the two corpora. IRT and IRPrT did not significantly differ in the rate of content word types and tokens used.

Table 8. Comparisons of measures of lexical density

Category	Indices	*M* (*SD*)		Significance Testing
		IRT	IRPrT	
Lexical density	types	.766 (.026)	.762 (.025)	t = 1.098, p = .27, d = .16
	tokens	.502 (.032)	.500 (.032)	t = .537, p = .59, d = .06

Table 9 shows the results of the comparisons of lexical diversity between the two corpora. No significant differences were observed in any of the measures of TTR, MATTR, MTLD, MTLD-MA-Wrap between IRT and IRPrT.

Table 9. Comparisons of measures of lexical diversity

Category	Indices	*M* (*SD*)		Significance Testing
		IRT	IRPrT	
TTR	aw	.430 (.031)	.434 (.034)	t = -1.509, p = .29, d = .12
	cw	104.81 (.058)	112.19 (.068)	z = -.868, p = .39, rs = .06
	fw	.202 (.020)	.206 (.019)	t = -1.669, p = .10, d = .20
MATTR	aw	.793 (.026)	.794 (.023)	t = -.280, p = .78, d = .04
	cw	106.47 (.033)	110.53 (.037)	z = -.477, p = .63, rs = .03
	fw	.517 (.041)	.519 (.037)	t = -.398, p = .69, d = .05
MTLD	aw	109.75 (16.552)	107.25 (15.862)	z = -.295, p = .77, rs = .02
	cw	106.39 (98.404)	110.61 (121.152)	z = -.496, p = .62, rs = .03
	fw	104.39 (2.996)	112.07 (2.613)	z = -.839, p = .40, rs = .06
MTLD-MA-Wrap	aw	109.59 (16.771)	107.41 (15.628)	z = -.257, p = .80, rs = .02
	cw	106.62 (89.892)	110.38 (104.612)	z = -.443, p = .66, rs = .03
	fw	104.68 (2.885)	112.32 (2.672)	z = -.898, p = .37, rs = .06

Syntactic Text Complexity

Table 10 shows the comparisons of syntactic complexity between the two corpora. The IRT exhibited greater length of t-units (MLT) with more verb phrases per t-unit than IRPrT, with both showing small effect sizes (*rs* = 0.15 and *rs* = 0.18, respectively), making these results somewhat inconclusive. Furthermore, the IRT exhibited more subordination than the IRPrT for all measuers with small effect sizes: C/T, *rs* = 0.15; CT/T, *d* = 0.33; DC/C, *d* = 0.30; DC/T, *rs* = 0.16.

Table 10. Comparisons of measures of syntactic complexity

Category	Indices	*M* (*SD*)		Significance Testing
		IRT	IRPrT	
Length of production unit	MLC	109.03 (1.944)	107.97 (1.844)	*z* = -.124, *p* = .90, *rs* = .01
	MLS	23.459 (3.010)	22.852 (3.284)	*t* = 1.416, *p* = .16, *d* = .20
	MLT	118.15 (2.914)	98.85 (3.052)	*z* = -2.270, *p* = **.02***, *rs* = .15
Sentence complexity	C/S	115.52 (.295)	101.48 (.283)	*z* = -1.652, *p* = .10, *rs* = .11
Subordination	C/T	117.60 (.268)	99.40 (.261)	*z* = -2.139, *p* = **.03***, *rs* = .15
	CT/T	.517 (.125)	.478 (.114)	*t* = 2.377, *p* = **.02***, *d* = .33
	DC/C	.404 (.077)	.381 (.077)	*t* = 2.190, *p* = **.03***, *d* = .30
	DC/T	118.25 (.236)	98.75 (.228)	*z* = -2.292, *p* = **.02***, *rs* = .16
Coordination	CP/C	102.29 (.143)	114.71 (.119)	*z* = -1.460, *p* = .14, *rs* = .10
	CP/T	105.09 (.227)	111.91 (.185)	*z* = -.802, *p* = .28, *rs* = .05
	T/S	103.90 (.077)	113.10 (.085)	*z* = -1.082, *p* = .28, *rs* = .07
Particular structures	CN/C	109.11 (.319)	107.89 (.352)	*z* = -.144, *p* = .89, *rs* = .01
	CN/T	116.01 (.540)	100.99 (.598)	*z* = -1.766, *p* = .08, *rs* = .12
	VP/T	119.71 (.399)	97.29 (.367)	*z* = -2.637, *p* = **.01***, *rs* = .18

** Significant difference at 0.05 level (2-tailed)*

Table 11 shows the comparisons of clausal and phrasal complexity between the two corpora. In terms of clausal complexity, the IRT exhibited greater number of direct objects per clause than the IRPrT with the small effect size (*d* = 0.30). In terms of phrasal complexity, the IRPrT exhibited more dependents per nominal subject than the IRT with the small effect size (*rs* = 0.19). However, the IRT exhibited more relative clause modifiers per nominal phrases than the IRPrT with the small effect size (*rs* = 0.13).

Table 11. Comparisons of measures of clausal and phrasal complexity

Indices	*M* (*SD*)		Significance Testing
	IRT	IRPrT	
Clausal complexity			
acomp	104.60 (.030)	112.40 (.037)	z = -.918, p = .36, rs = .06
advcl	106.35 (.024)	110.65 (.028)	z = -.505, p = .61, rs = .03
advmod	.214 (.052)	.210 (.058)	t = .477, p = .63, d = .08
ccomp	115.91 (.050)	101.09 (.047)	z = -1.742, p = .08, rs = .12
cc	107.50 (.014)	109.50 (.018)	z = -.236, p = .81, rs = .02
conj	106.20 (.040)	110.80 (.037)	z = -.541, p = .59, rs = .04
mark	100.38 (.045)	116.63 (.045)	z = -1.911, p = .06, rs = .13
pcomp	105.00 (.002)	112.00 (.004)	z = -1.765, p = .08, rs = .12
csubj	107.82 (.011)	109.18 (.009)	z = -.169, p = .87, rs = .01
xsubj	N/A	N/A	N/A
nsubj	.609 (.089)	.602 (.082)	t = .565, p = .57, d = .08
dobj	.400 (.072)	.378 (.075)	t = 2.225, p = **.03***, d = .30
iobj	105.92 (.005)	111.08 (.006)	z = -.784, p = .43, rs = .05
ncomp	103.03 (.031)	113.97 (.035)	z = -1.286, p = .20, rs = .09
xcomp	114.79 (.032)	102.21 (.035)	z = -1.479, p = .14, rs = .10
Phrasal complexity			
nsubj_deps	96.86 (.200)	120.14 (.244)	z = -2.738, p = **.01***, rs = .19
ncomp_deps	110.17 (.828)	106.83 (.803)	z = -.393, p = .69, rs = .03
dobj_deps	107.25 (.235)	109.75 (.240)	z = -.294, p = .77, rs = .02
iobj_deps	107.37 (.218)	109.63 (.418)	z = -.585, p = .56, rs = .04
pobj_deps	1.343 (.126)	1.313 (.125)	t = 1.735, p = .08, d = .24
det_nominal	.312 (.064)	.315 (.053)	t = -.301, p = .76, d = .05
amod_nominal	103.14 (.050)	113.86 (.058)	z = -1.261, p = .21, rs = .09
prep_nominal	.220 (.044)	.218 (.049)	t = .335, p = .74, d = .05
poss_nominal	115.39 (.023)	101.61 (.025)	z = -1.621, p = .11, rs = .11
vmod_nominal	106.17 (.014)	110.83 (.014)	z = -.549, p = .58, rs = .04
nn_nominal	106.81 (.056)	110.19 (.055)	z = -.398, p = .69, rs = .03
rcmod_nominal	116.75 (.014)	100.25 (.014)	z = -1.941, p = **.05***, rs = .13
advmod_nominal	103.84 (.013)	113.16 (.121)	z = -1.096, p = .27, rs = .07
conj_and_nominal	101.63 (.025)	115.37 (.026)	z = -1.615, p = .11, rs = .11
conj_or_nominal	103.37 (.008)	113.63 (.010)	z = -1.221, p = .22, rs = .08

** Significant difference at 0.05 level (2-tailed)*

Discourse Text Complexity

Table 12 shows the comparisons of the discourse level indices of the IRT and IRPrT. The IRT had higher lexical overlap than IRPrT in all indices at both sentence and paragraph levels with the large effect sizes. This indicates that the IRT had more repetition of words, including nouns and pronouns in subsequent sentences than the IRPrT. The IRT also had much more semantic overlap than the IRPrT, as all of the indices exhibited significant differences with the large effect sizes. This means that the IRT had much more semantic similarity between all adjacent sentences, and paragraphs than the IRPrT.

Finally, the results also show that the IRPrT employed more basic connectives, demonstratives, additives and logical connectors than IRT with the large effect sizes. Furthermore, the IRPrT had much higher givenness indices than the IRT as indicated by the large effect sizes. This suggests that content words and third person pronouns were repeated far more in the IRPrT than in the IRT.

Table 12. Comparisons of measures of discourse text complexity

Indices	*M* (*SD*)		Significance Testing
	IRT	IRPrT	
Lexical Overlap			
2_argument_sent	161.98 (.041)	55.02 (.028)	*z* = -12.557, *p* = **.00***, *rs* = .85
binary_all_sent	162.50 (.053)	54.50 (.058)	*z* = -12.705, *p* = **.00***, *rs* = .86
binary_argument_sent	162.48 (.131)	54.52 (.047)	*z* = -12.695, *p* = **.00***, *rs* = .86
2_all_para	161.63 (.058)	55.38 (.018)	*z* = -12.508, *p* = **.00***, *rs* = .85
2_argument_para	159.81 (.063)	57.19 (.037)	*z* = -12.065, *p* = **.00***, *rs* = .82
binary_all_para	162.50 (.065)	54.50 (.045)	*z* = -13.120, *p* = **.00***, *rs* = .89
binary_argument_para	162.50 (.135)	54.50 (.004)	*z* = -13.399, *p* = **.00***, *rs* = .91
Semantic overlap			
lsa_1_all_sent	162.50 (.071)	54.50 (.009)	*z* = -12.780, *p* = **.00***, *rs* = .87
lsa_2_all_sent	162.50 (.047)	54.50 (.000)	*z* = -13.575, *p* = **.00***, *rs* = .92
lsa_1_all_para	146.13 (.082)	70.88 (.119)	*z* = -8.848, *p* = **.00***, *rs* = .60
lsa_2_all_para	160.81 (.109)	56.19 (.075)	*z* = -12.301, *p* = **.00***, *rs* = .84
Connectives			
basic_connectives	54.56 (.006)	162.44 (.008)	*z* = -12.938, *p* = **.00***, *rs* = .88
all_demonstratives	63.66 (.007)	153.34 (.035)	*z* = -10.546, *p* = **.00***, *rs* = .72
all_additive	64.95 (.010)	152.05 (.035)	*z* = -10.242, *p* = **.00***, *rs* = .70
all_logical	54.50 (.009)	162.50 (.244)	*z* = -12.701, *p* = **.00***, *rs* = .86
Givenness			
repeated_content_lemma	56.50 (.039)	160.50 (.803)	*z* = -12.230, *p* = **.00***, *rs* = .83
repeated_content_and_pronoun_lemma	54.50 (.039)	162.50 (.240)	*z* = -12.699, *p* = **.00***, *rs* = .86

** Significant difference at 0.05 level (2-tailed)*

Discussion

Lexical Text Complexity

The findings indicate that the IRT and IRPrT did not significantly differ in lexical sophistication. Specifically, the IRT and IRPrT had relatively equal text coverage of the most frequent 3,000 and 5,000 words; around 94% of texts in both being covered at the 3,000-word level, and approximately 97% being covered at the 5,000-word level. This suggests that the vocabulary loads in the IRT and IRPrT are quite similar. Moreover, no significant difference was shown in word frequency, word range, and contextual distinctiveness between IRT and IRPrT. This means that lexical items, either content or function words, are encountered a similar number of times across similar contexts in both corpora. This suggests that the IRPrT reflect the IRT well in terms of vocabulary size; therefore, if learners develop vocabulary knowledge through the IRPrT, they are likely to encounter words of the same frequency levels in the IRT.

The findings also indicate that academic language was used at relatively equal frequency in the IRT and IRPrT. Therefore, the IRPrT are also likely to prepare learners for academic language that will appear on the IRT. However, it should be noted that the AWL coverage in both corpora was far lower than previous studies of academic written English texts (Chen & Ge, 2007; Vongpumivitch et al., 2009). This raises the question as to whether the IRT is truly a valid measure of the academic vocabulary knowledge learners will need in real-life language domains later. However, future studies would be warranted to know this for sure.

We also found no difference in the lexical sophistication of the IRT and IRPrT. Both word recognition norms (i.e., the time amount required to identify words correctly and to name them) and age of exposure (i.e., no words in any of the corpora required a more sophisticated link to other lexical items of relevant meaning for understanding the texts) measures were comparable

across the corpora. This suggests a similar balance in cognitive processing of vocabulary when reading texts from either.

Results for lexical density showed no difference between the IRT and IRPrT. Both IRT and IRPrT were comprised of almost 76% content words. This implies that the IRPrT expose learners to equal cognitive processing levels for semantically-rich words as the real test conditions, meaning that this sort of test preparation content is aligned quite well with the tests. Furthermore, the IRT and IRPrT had relatively equal density of content and function word tokens (approximately 50%). This further suggests that the IRPrT are providing a representative mix of content and function words, semantically tied together for comprehensibility.

The IRT and IRPrT did not significantly differ in lexical diversity indices of TTR, MATTR, MTLD, and MTLD-MA. Both corpora contained around 43% different words overall, 66% different content words, and 20% different function words. This indicates less repetition of content words than function words in both. The high lexical diversity level of content words in the IRPrT may help test takers when taking the IRT, as there is a similar amount of variety in the real tests.

Based on these results, the lexical alignment has positive implications for learners as follows:

- First, IRPrT can expose learners to texts with a similar informational load and mix of content and grammatical elements as found in IRT. This means that these materials provide truthful test practice at the word level and learners are likely developing relevant lexical knowledge through IRPrT.
- Second, as learners are aware that the words they learn in practice conditions can be of help in test conditions, vocabulary knowledge is more likely to be acquired and transferable to decoding meanings on IRT. This therefore implies positive washback - the practice tests motivate development of vocabulary knowledge construct-relevant to the actual test (Green, 2007). Seeing their word level improve on IRPrT may reinforce students' confidence to tackle the lexical complexity of IRT.

Additionally, the lexical alignment can bear useful implications for instructors and IRPrT material developers. As for instructors, they can be further assured that assigning IRPrT materials can aid learners' expansion of lexicons in high-frequency and academic tiers, thereby being reasonably assured of parity with distributions in high-stakes IRT. Similarly, equipped with quantitative evidence that IRPrT exhibit comparable lexical sophistication, materials developers can utilize said evidence to validate the appropriateness of lexical profile in future IRPrT based on IRT benchmarks. If future IRPrT align with the lexical profile documented here, positive washback transpires as the future IRPrT may equip learners' lexicons to meet the demands of IRT, hence reinforcing score validity.

Syntactic Text Complexity

Clear differences were found in the syntactic complexity of the IRT and IRPrT corpora. Specifically, there might be a small difference in the length of syntactic units between the tests, as indicated by the significant difference found in MLT. However, since there were no differences in MLC and MLS, and there are differences in how older versions of TAASSC and Lu's (2010) original L2SCA calculate T-units, the difference found in MLT is somewhat inconclusive. These results might be affected by the fact that we used an older version of TAASSC which calculates T-units based on dependency tags, which is different from the tregex method used by Lu (2010). More importantly, subordination, a syntactic device that contributes to the depth and intricacy of sentences, was found to be used more frequently in the

IRT than in the IRPrT across all measures of subordination (C/T, CT/T, DC/C and DC/T). For instance, the IRT exhibited a higher density of clauses per T-unit than the IRPrT. This pattern was echoed across the other subordination measures, i.e., the number of complex T-units divided by T-units as well as the numbers of dependent clauses per clause and T-unit. These findings indicate that the test uses far more of these complex structures than the practice materials, which is argued to be significant because potential test takers are likely to be under a more stressful cognitive load to decode the syntactic packaging as compared with practice conditions. For this reason, their reading comprehension might be impeded in the real-time processing of test conditions, as aligned with the previous studies (e.g., Kyle, 2016; Mesmer et al., 2012) where more syntactic complexity contributes to greater difficulty in reading comprehension. Moreover, because of no significant difference in lexical text complexity between the IRT and IRPrT as discussed earlier, greater syntactic complexity in the IRT may cause difficulty for those test takers having limited competence to negotiate with syntactic relations of the IRT in actual test conditions no matter how well-prepared they are for lexical knowledge in the IRPrT during test preparation. This interpretation corroborates Shiotsu and Weir's (2017) and Tong et al.'s (2024) studies where syntactic knowledge and skills has more effect on lexical knowledge in text processing and comprehension.

While our analyses included several indices of clausal and phrasal complexity, only a few indices mentioned above showed significant differences between the IRT and IRPrT. At the clausal level, a higher number of direct objects per clause in the IRT also means a more intricate description of the subject's actions for learners to interpret in the test. At the phrasal level, greater relative clause modifiers in the IRT can be mentally taxing to test takers in unpacking the meaning of sentences in the test condition. Although these isolated findings suggest caution in generalizing differences in clausal and phrasal complexity between the IRT and IRPrT, they may show that clausal and phrasal complexity is somewhat underrepresented in the IRPrT. Greater clausal embeddings to phrases and more direct objects per clauses can impose heavy cognitive loads on readers as well as impede processing time and comprehension. This is considered to be problematic, because of the limited capacity of one's working memory. "[U]nder normal conditions, information can be remembered in working memory for about two seconds only. After that brief span, the representation is rapidly forgotten, unless it can be rehearsed subvocally" (Ortega, 2009, p. 90). Based on this premise, if test takers encounter texts with greater clausal and phrasal complexity under the time constraint of the test condition than in the practice conditions, they may be under cognitive pressure to commit the information they read to the site of consciousness (Baars & Franklin, 2003).

However, the mean for dependents per nominal subject is higher in the IRPrT than the IRT. Given that a sentence with numerous dependents may be considered redundant (Crossley et al., 2017), we would argue that this difference could be problematic in terms of syntactic complexity. As redundant information could lack conciseness and hinder the effective communication of ideas, this may unnecessarily overload learners with a false impression about what to expect in the IRT and may arouse anxiety as a form of negative washback during test preparation (Nguyen, 2023).

One pedagogical implication of these results is that classroom instructors can leverage these findings to better prepare learners for the sophisticated structures of IRT under strict time constraints. Making students aware of the syntactic gap could motivate them to seek additional practice to regulate learning and expand readiness for syntactic complexity in IRT. Furthermore, supplementary materials or instruction focused on comprehending longer T-units and clauses with multiple embedded elements may be warranted. Through targeted practice on the enhanced materials, students can develop strategies to unpack meaning from sophisticated syntax they are likely to encounter on IRT and dispel misconceptions about upcoming IRT.

Such a coordination between instruction and further material use guided by empirical syntactic data serves to reduce negative washback during test preparation as the threat of construct-irrelevant variance to test scores. At the same time, future material developers of IRPrT could consider calibrating the syntax of IRPrT more closely to the sophistication of IRT.

Discourse Text Complexity

We found that the IRT had higher overlap than the IRPrT at both sentence and paragraph levels. This finding suggests that the IRPrT are not as lexically or semantically cohesive as the IRT, which might make them more difficult to comprehend (Gernsbacher, 2013). We would argue that having to read more challenging texts without lexical cohesion from test preparation materials at hand, learners may exhibit disengagement and reduced confidence when preparing for the IELTS.

The semantic overlap results indicated that there was greater semantic similarity in the IRT than in the IRPrT between all adjacent sentences (both at one-sentence and two-sentence intervals) and between all adjacent paragraphs (both at one-paragraph and two-paragraph intervals). Therefore, it can be inferred that there was less similarity of ideas in the IRPrT than the IRT. Given that local (sentence) cohesion of ideas can facilitate moment-to-moment understanding of a text and that global (text) cohesion can be conducive to the overall integration and recall of ideas as the text progresses (Koda, 2005), the lack of local and global cohesion in the IRPrT can be linked to difficulty in reading comprehension during test preparation, as suggested by Ehrlich (1991).

There are several potential reasons for the greater lexical and semantic overlap in the IRT than the IRPrT. One is that trade-off between cohesion and syntactic sophistication increased text cohesion but led to greater text length, density, and complexity (Beck et al., 1991; Ozuru et al., 2009). This could potentially also explain why MLT seemed to be slightly longer in IRT than in IRPrT. Another reason could be due to the fact that the IRPrT had more basic connectives, demonstratives, additives and logical connectors than IRT. Although having more givenness and connectives would support learners in linking ideas of a reading text easily to a certain extent during test preparation, we would argue that caution should be exercised in generalizing these findings to be a complete advantage for learners. Without a blend of semantic similarity at different discourse distances in IRPrT, learners may struggle to temporally commit mental representations of text information to their working memory in line with the text development. The disjointed flow of meaningful ideas during reading may lead them to lose focus on a text at hand and score low on reading comprehension items during test preparation, so IRPrT materials may consider reducing these connectors in favor of repetition in the future.

Based on these findings, some implications can be drawn for classroom instructor, material developers and learners as follows:

- Classroom instructors should take greater care for selecting preparatory materials that cultivate the cohesive flow found in IRPrT, not just the usage of connectives. They should also be attentive that relying excessively on transitional phrases in lesson texts may not be conducive to the inherent lexical and semantic ties. Classroom exercises and readings should also provide exposure to and emphasize the importance of lexical overlap, phrasal repetition, and semantic connectivity threaded throughout texts, modeling the cohesion patterns on IRT.
- Materials developers should rigorously analyze the discourse-level cohesion of IRPrT by developing the passages exhibiting phrasal repetition and semantic similarity between sentences and paragraphs on par with IRT. Conscious efforts made to mirror

the authentic cohesive flow of the IRT itself, not merely with focus on connective usage, will better prepare learners for the linguistic demands of IRT.

- Learners should not fall to the misconception that just having more connecting words means having better discourse-level cohesion. When engaging with IRPrT, learners should consciously observe how coherent flow arises in passages from lexical and phrasal repetition and tight semantic ties, not solely transitional words.

This study also bears broader implications beyond IELTS for language testing researchers by offering valuable operationalization for investigating other high-stakes ELPT used for admission to academic contexts. A parallel multidimensional complexity audit comparing texts from such widely-used ELPT as TOEFL, IELTS, and PTE academic in relation to their corresponding practice tests would enable detailed comparability in linguistic features. Such empirical profiling illuminates if the text complexity features appropriately align across these standardized tests themselves as well as between the real tests and their preparatory materials. Given that English language testing increasingly plays a critical gatekeeping role for high-stakes decisions, an understanding of comparative text complexity features across the key ELPT can strengthen claims of score equivalence, and validity of score interpretations. Another implication pertains to profiling innovative informal assessments, such as the online Duolingo English Test, which have claimed increased acceptance for tertiary admissions. A systematic analysis of Duolingo's linguistic complexity in comparison with other traditional ELPT would determine areas of alignment and divergence across lexical, grammatical and discourse features. This provides a nuanced perspective on the relationships emerging between established standards and "informal" online formats that promise greater access yet require empirical profiling of construct coverage, difficulty, cognitive processing and comparability to traditional benchmarks. In totality, by undertaking corpus-based multidimensional text complexity analysis, this study offers research-driven implications beyond merely IELTS preparation to examining the intricate relationships between traditional and technology-mediated ELPT, buttressing consequential decisions.

While this study offers valuable insights, certain limitations should be noted. First, the analysis solely employed computational tools to quantify text complexity differences. The subjective experience of learner readers was not incorporated. Perception studies measuring learners' comprehension, engagement, and strategies in IRT versus IRPrT could complement the corpus-based computational comparisons. To further illuminate the comparability of text complexity between IRT and IRPrT, future studies could compare test-takers' reading performance on passages drawn from the two corpora. Simulated recall procedures made immediately after reading could further shed light on how the readers navigated and interpreted linguistic complexity variations between IRPrT and IRT. Second, the links between text complexity measures and actual reading performance remain unexplored. Relating the text complexity measures to test-takers' scores at different proficiency levels could better explain the implications. Despite such limitations, the study is one of the few corpus-based investigations to compare IRPrT with IRT in terms of text complexity and provide evidence on IRPrT usefulness in preparation for IRT.

Conclusions

The results of this study show that the IRT and IRPrT are genreally quite comparable in terms of objective measures of lexical text complexity, indicating that the IRPrT can provide practice that is likely to be lexically similar to IRT and prepare test takers for the real IELTS. This is important as it assures educators and learners that the IRPrT are providing authentic lexical practice. However, the results also indicate that the IRT might have slightly longer syntactic

units and that it definitely uses much more subordination than IRPrT. Therefore, test takers using the IRPrT might want to be aware of this difference and supplement their learning with some other materials that help with reading sentences with lots of suboordination. Finally, we found that the IRT has a larger amount of cohesion, specifically semantic repetition, than the IRPrT texts, which instead favor connectors. Therefore, educators and learners using these materials might want to practice reading texts with more repetition, and IRPrT materials developers may want to consider increasing semantic repetition in future practice texts that they create.

About the Authors

Huu Thanh Minh Nguyen is a lecturer in English as a Foreign Language at University of Foreign Language Studies, The University of Danang. He earned his Master's degree in TESOL Studies from the University of Leeds, UK. He is doing doctoral research at School of Languages and Linguistics, The University of Melbourne, Australia. His main research interests include Language Testing and Assessment, Second Language Writing, and Corpus Linguistics. ORCID ID: 0000-0002-9494-9641.

Nguyen Van Anh Le is a lecturer in English as a Foreign Language at University of Foreign Language Studies, The University of Danang. She earned her Master's degree from the University of Huddersfield, UK, and her Ph.D. from Sophia University, Tokyo, Japan. Her main research interests include Phonetics, Phonology, Second Language Acquisition, and Sociolinguistics. ORCID ID: 0009-0000-7378-1432

Acknowledgements and Funding

We owe our immense thanks to the editorial board and four anonymous reviewers whose insightful and critical reviews have contributed to critical improvements in the earlier versions of the manuscript. We would also want to extend our deepest gratitude to Mr. Ngo Pham Anh Huy whose coding support was an indispensable part of our work.

Research for this article was funded by University of Foreign Language Studies, The University of Danang (Project number: T2023-05-15) and Funds for Science and Technology Development, The University of Danang.

Note

[1] T-unit is known as one main clause plus any subordinate clause or non-clausal structure to which it is attached or within which it is embedded (Vajjala & Meurers, 2013).

[2] A lemma is a "dictionary headword; an abstract representation, subsuming all the formal lexical variations which may apply: the verb *walk*, for example, subsumes *walking*, *walks* and *walked*" (Knowles & Mohd Don, 2004, p. 70, original emphasis).

[3] Givenness indices are reflected of the approximated proportion of given information to new information (Crossley et al., 2016).

To Cite this Article

Nguyen, H. T. M., & Le, N. V. A. (2024). Text Complexity of Cambridge-delivered IELTS Academic Reading Tests: Comparability with IELTS Academic Reading Practice Tests from Other Publishers. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 28*(2). https://doi.org/10.55593/ej.28110a4

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Baars, B. J., & Franklin, S. (2003) How conscious experience and working memory interact. *TRENDS in Cognitive Sciences*, *7*(4), 166–172. https://doi.org/10.1016/S1364-6613(03)00056-1

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/10.1177/026553229601300202

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning and use*. Blackwell.

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: evidence of improved comprehensibility. *Reading Research Quarterly*, *26*(3), 251–276. https://doi.org/10.2307/747763

Biber, D., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond t-units-a response to Weiwei Yang. *TESOL Quarterly*, *47*(1), 192–201. https://doi.org/10.1002/tesq.84

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, *20*, 7–36.

Briscoe, T., Medlock, B., & Andersen, Ø. E. (2010). *Automated assessment of ESOL free text examinations*. University of Cambridge: Computer Laboratory. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-790.pdf

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. House, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Carrell, P. L. (1987). Readability in ESL. *Reading in a Foreign Language*, *4*(1), 21–40.

Chen, Q. & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*, 502–504. https://doi.org/10.1016/j.esp.2007.04.003

Cobb, T. (2009). *The Compleat lexical tutor*. http://www.lextutor.ca/.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

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

Ehrlich, M. F. (1991). The processing of cohesion devices in text comprehension. *Psychological Research, 53*(2), 169–174. https://doi.org/10.1007/BF01371825

Everett, R., Coleman, J. (2003, April 17). *A critical analysis of selected IELTS preparation materials.* IELTS. https://ielts.org/researchers/our-research/research-reports/a-critical-analysis-of-selected-ielts-preparation-materials

Fang, Z., & Pace, B. G. (2013). Teaching with challenging texts in the disciplines: Text complexity and close reading. *Journal of Adolescent & Adult Literacy*, *57*(2), 104–108. https://doi.org/10.1002/JAAL.229

Fulmer, S. M., D'Mello, S. K., Strain, A., & Graesser, A. C. (2015). Interest-based text preference moderates the effect of text difficulty on engagement and learning. *Contemporary Educational Psychology*, *41*, 98–100. https://doi.org/10.1016/j.cedpsych.2014.12.005.

Garnier, M. & Schmitt, N. (2015). The PHaVE List: A pedagogical list of PVs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 645–666. https://doi.org/10.1177/1362168814559798

Gernsbacher, M. A. (2013). *Language comprehension as structure building*. Psychology Press.

Goldman, S. R., & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Rosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research*, (pp. 311–335). Lawrence Erlbaum.

Gough, C., & Hutchison. S. (2015). *Exam Essential Practice Tests: IELTS 2*. Cengage Learning.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. http://doi.org/10.3758/BF03195564

Green, A. (2006). Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes, *Assessing Writing, 11*(2), 113–134. https://doi.org/10.1016/j.asw.2006.07.002

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

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Unpublished doctoral dissertation. Temple University.

Guthrie, J. T., Klauda, S. L., & Ho, A. N. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Reading Research Quarterly*, *48*(1), 9–26. https://doi.org/10.1002/rrq.035

Halliday, M. A. K. (2004). *An introduction to functional grammar*. Hodder Education.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Harrison, M., & Whitehead, R. (2015). *Exam Essential Practice Tests: IELTS 1*. Cengage Learning.

Hoeksema, J., & Napoli, D. J. (1993). Paratactic and subordinative *so*. *Journal of Linguistics, 29*(2), 291–314. https://doi.org/10.1017/s0022226700000347

Jakeman, V., & McDowell, C. (2001). *IELTS Practice Tests Plus*. Pearson.

Kirby, P. (2016). *Shadow schooling: Private tuition and social mobility in the UK*. The Sutton Trust.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Knowles, G., & Mohd Don, Z. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, *9*(1), 69–81. https://doi.org/10.1075/ijcl.9.1.04kno

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* Cambridge University Press.

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

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.* Doctoral dissertation, Georgia State University. https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*. Advance online publication. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learner's vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Matthews, M., & Salisbury, K. (2011). *IELTS Practice Tests Plus 3*. Pearson.

McCarter, S., & Ash, J. (2008). *IELTS Test Builder 1*. Macmillan.

McCarter, S. (2008). *IELTS Test Builder 2*. Macmillan.

McEnery, T. (2006). A2. Representativeness, balance and sampling. In T. McEnery, R. Xiao, & Y. Tono (Eds.), *Corpus-based language studies: An advanced resource book* (pp. 13–21). Routledge.

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*(3), 235–258. https://doi.org/10.1002/rrq.019

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568-587. https://doi.org/10.1016/j.jml.2012.03.008

Messick, S. (1989). Validity. In R. L. Linn, (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). American Council on Education and Macmillan.

Michel, M. (2017). Complexity, Accuracy and Fluency (CAF). In: L. Shawn, & S. Masatoshi (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (pp. 50–68). Routledge.

Nation, I. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nguyen, H. T. M. (2023). The washback of the International English Language Testing System (IELTS) as an English language proficiency exit test on the learning of final-year English majors. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 27*(2), 1–34. https://doi.org/10.55593/ej.27106a8

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518. http://doi.org/10.1093/applin/24.4.492

Ortega, L. (2009). *Understanding Second language acquisition*. Hodder Education.

O'Sullivan, B., Dunn, K., & Berry, V. (2021). Test preparation: an international comparison of test takers' preferences. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 13–36. https://doi.org/10.1080/0969594X.2019.1637820

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal, 73*(2), 197–206. https://doi.org/10.1093/elt/ccz006

Perfetti C., & Stafura J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Sainani, K. L. (2012). Dealing with Non-normal Data. *PM&R*, *4*(12), 1001–1005. https://doi.org/10.1016/j.pmrj.2012.10.013

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128. https://doi.org/10.1177%2F0265532207071513

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Education.

Spring, R. (2019). *From Linguistic Theory to the Classroom: A Practical Guide and Case Study*. Cambridge Scholars.

Terry, M., & Wilson, J. (2005). *IELTS Practice Tests Plus 2*. Pearson.

Thompson, G. (2014). *Introducing functional grammar*. Routledge.

Tong, X., Yu, L., & Deacon, S. H. (2024). A Meta-Analysis of the Relation Between Syntactic Skills and Reading Comprehension: A Cross-Linguistic and Developmental Investigation. *Review of Educational Research*. https://doi.org/10.3102/00346543241228185

Vajjala, S., & Meurers, D. (2013). On the applicability of readability models to web texts. In S. Williams, A. Siddharthan, & A. Nenkova (Eds.), *Proceedings of the 2nd workshop on predicting and improving text readability for target reader populations* (pp. 59–68). Association for Computational Linguistics. https://aclanthology.org/W13-2907.pdf

Vongpumivitch, V., Huang J., & Chang Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*, 33–41. https://doi.org/10.1016/j.esp.2007.04.003

Yang, Y.-H., Chu, H.-C., & Tseng, W.-T. (2021). Text Difficulty in Extensive Reading: Reading Comprehension and Reading Motivation. *Reading in a Foreign Language*, *33*(1), 78–102.

Yu, X. (2021). Text Complexity of Reading Comprehension Passages in the National Matriculation English Test in China: The Development from 1996 to 2020. *International Journal of Language Testing*, *11*(2), 142–167.