IELTS Research Reports Online Series

ISSN 2201-2982
Reference: 2016/3

What changes and what doesn't?
An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts

Authors: *Khaled Barkaoui, Faculty of Education, York University, Toronto, Canada*

Grant awarded: *Round 19, 2013*

Keywords: *"IELTS Writing Task 2, linguistic characteristics, test repeaters, multilevel modelling, longitudinal study, second-language writing"*

Abstract

This study examined changes in the linguistic characteristics of IELTS repeaters' responses to Writing Task 2. It analysed 234 scripts written by 78 candidates who belonged to three groups in terms of their initial writing abilities. The candidates each took IELTS Academic three times.

Various computer programs were used to analyse the scripts in terms of features related to the candidates':

- grammatical choices, i.e., fluency, accuracy, syntactic complexity and lexical features
- discourse choices, i.e., coherence and cohesion, discourse structure
- sociolinguistic choices, i.e., register
- strategic choices, i.e., interactional metadiscourse markers.

The findings indicated that scripts with higher writing scores at test occasion 1 were more likely to include an introduction and a conclusion and tended to be significantly longer, to have greater linguistic accuracy, syntactic complexity, lexical density, diversity and sophistication, and cohesion, and to include longer introductions and conclusions, fewer informal features (i.e., contractions), more formal features (i.e., passivisation, nominalisation), more hedges, and fewer self-mentions than did scripts with lower writing scores.

Generally, scripts produced at later test occasions tended to be significantly longer, more linguistically accurate, more coherent, and to include more formal features (i.e., passive constructions and nominalisation) and fewer interactional metadiscourse markers than scripts produced at earlier test occasions.

While the rate of change over time for some of these features (e.g., fluency, nominalisations) varied significantly across candidates, initial L2 writing ability did not significantly moderate the rate of change in these features.

Finally, longer scripts with greater lexical diversity and lexical sophistication, greater syntactic complexity, more self-mentions, and fewer contractions tended to obtain higher writing scores.

The findings of the study are consistent with previous studies on IELTS Writing Task 2, but they also highlight the value of examining repeaters' test performance and point to several areas for further research.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2016.
This online series succeeds *IELTS Research Reports Volumes 1–13*, published 1998–2012 in print and on CD.

AUTHOR BIODATA

Khaled Barkaoui

Khaled Barkaoui is an Associate Professor at the Faculty of Education, York University, Canada. His current research and teaching focus on second-language (L2) assessment, L2 writing, L2 program evaluation, longitudinal and mixed-methods research, and English for Academic Purposes (EAP). His publications have appeared in *Applied Linguistics, Assessing Writing, Language Testing, Language Assessment Quarterly, System* and *TESOL Quarterly*.

In 2012, Khaled received the *TOEFL Outstanding Young Scholar Award* in recognition of the outstanding contributions his scholarship and professional activities have made to the field of second language assessment.

ACKNOWLEDGEMENTS

I would like to thank the following people:

- the IELTS partners for funding this study
- Ibtissem Knouzi for helping with the preparation of the scripts for computer analyses and for conducting the computer analyses of the scripts
- Shouzheng Li for helping with data organisation, preparation and entry for statistical analyses
- Amy Lee for editing an earlier draft of the report for style.

The opinions expressed in the report are those of the author and do not necessarily reflect the views of the IELTS partners.

IELTS Research Program

The IELTS partners – British Council, Cambridge English Language Assessment and IDP: IELTS Australia – have a longstanding commitment to remain at the forefront of developments in English language testing.

The steady evolution of IELTS is in parallel with advances in applied linguistics, language pedagogy, language assessment and technology. This ensures the ongoing validity, reliability, positive impact and practicality of the test. Adherence to these four qualities is supported by two streams of research: internal and external.

Internal research activities are managed by Cambridge English Language Assessment's Research and Validation unit. The Research and Validation unit brings together specialists in testing and assessment, statistical analysis and item-banking, applied linguistics, corpus linguistics, and language learning/pedagogy, and provides rigorous quality assurance for the IELTS test at every stage of development.

External research is conducted by independent researchers via the joint research program, funded by IDP: IELTS Australia and British Council, and supported by Cambridge English Language Assessment.

Call for research proposals

The annual call for research proposals is widely publicised in March, with applications due by 30 June each year. A Joint Research Committee, comprising representatives of the IELTS partners, agrees on research priorities and oversees the allocations of research grants for external research.

Reports are peer reviewed

IELTS Research Reports submitted by external researchers are peer reviewed prior to publication.

All IELTS Research Reports available online

This extensive body of research is available for download from www.ielts.org/researchers.

INTRODUCTION FROM IELTS

This study by Khaled Barkaoui of York University in Canada was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia and Cambridge English Language Assessment) as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with more than100 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in several IELTS-focused volumes in the *Studies in Language Testing* series (www.cambridgeenglish.org/silt), in academic journals, and in *IELTS Research Reports*. Since 2012, in order to facilitate timely access, individual research reports have been made available on the IELTS website immediately after completing the peer review and revision process.

This report looks at the writing of IELTS Academic candidates at various ability levels, and the way their writing changes on multiple subsequent sittings of the test. Unlike earlier studies (e.g. Elder & O'Loughlin, 2003; Green, 2005), this study is an attempt at a longitudinal view of repeat candidates' performance on the test. To do this, the study employs multilevel modelling, which has been around for a while in education research, but is only now making its way into language assessment research, primarily through Barkaoui's efforts.

To explain briefly: In education research, regression techniques have been a central tool for performing quantitative analysis. However, an assumption of these techniques is that observations are independent of one another. This is often not the case with education data. For example, students' scores are not independent because they are a function of the classrooms that they belong to and teachers they have been taught by. Multilevel modelling (MLM) is an approach which takes into account such 'nested' data. A happy consequence of MLM is that repeated measures (such as with repeat IELTS candidate scores) can be seen as nested within particular candidates. That is, the method can be used to investigate longitudinal data.

Admittedly, this is a relatively modest attempt at that, as the data was limited to three observations per candidate, and the analysis did not try to account for the different amounts of time that had elapsed between test sittings for different candidates, which future research can and should address. Nonetheless, the picture presented is of candidates improving, not just in band score terms, but also in certain measurable features of their writing.

I say measurable features because, while many computational tools have been developed of late to quantify text features (e.g. Coh-Metrix, AntConc), as the report acknowledges, there remain valued qualities of good writing that do not easily lend themselves to quantification. Indeed, it may be that some of these qualities disappear from view when texts are broken down into smaller and smaller units.

In sum, the report makes a contribution by demonstrating how one tool can be useful in the conduct of language assessment research, even as it shows the limitations of some of our other tools. For language assessment research, a frontier has been crossed, but more frontiers beckon in the horizon.

Dr Gad S Lim, Principal Research Manager
Cambridge English Language Assessment

References to the IELTS Introduction

Elder, C. and O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gain on IELTS. *IELTS Research Reports*, *Vol 4*, R. Tulloh (Ed.), IELTS Australia Pty Limited, Canberra, pp. 5–43.

Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test. *Assessing Writing*, *10*, pp. 44–60.

CONTENTS

List of tables

This study aimed to examine *changes over time* in the linguistic characteristics of texts written in response to IELTS Writing Task 2 by candidates who took IELTS Academic three times. The valid interpretation and use of second-language (L2) test scores rests on several assumptions, including the assumption that test scores vary depending on candidates' L2 proficiency as demonstrated in their test performance (Chapelle, 2008; Weir, 2005). In a L2 writing test, this means that test scores vary as a function of the quality of candidates' texts, which in turn vary in relation to candidates' L2 proficiency. More proficient candidates are expected to produce better-quality texts (e.g., texts with fewer errors, better coherence), which will receive higher scores than will poorer-quality texts produced by less proficient candidates. The typical approach to examine this assumption is to conduct a cross-sectional study that compares the linguistic characteristics of scripts at different score levels at one time point (e.g., Banerjee et al., 2007; Cumming et al., 2005; Riazi and Knox, 2013). Evidence supporting the assumption above strengthens the validity of score-based inferences about candidates' L2 writing abilities.

This validity question can also be addressed using a longitudinal design to examine the relationship between *changes* in the writing features of the scripts of *the same candidates* and changes in their writing scores *over time*. This can be achieved by comparing the scripts and test scores of test repeaters across testing occasions. A key assumption underlining the interpretation of repeaters' writing scores is that changes in their writing test scores reflect true changes in relevant linguistic characteristics of their texts over time. Another assumption is that changes in the characteristics of repeater' texts, in turn, reflect true changes in their L2 writing abilities over time. To the extent that empirical evidence backs both assumptions, the test's validity argument is supported. The following sections review previous research on test repeaters and the writing features that distinguish scripts at different L2 proficiency levels.

1 BACKGROUND

1.1 Previous studies on test repeaters

A central question in test validation research concerns the meaning of test scores. This question is often investigated by examining factors that contribute to variability in test scores at one point in time. Few studies have investigated this question longitudinally by examining score changes across time. Most of these studies were done in relation to IELTS and fall into two categories (Green, 2005): (a) studies that compared the scores of candidates who took the test twice (e.g., Green 2005) and (b) studies that compared the scores of L2 learners who took the test before and after relevant English language instruction (e.g., Brown, 1998; Elder and O'Loughlin, 2003; O'Loughlin and Arkoudis, 2009; Rao et al., 2003; Read and Hayes, 2003). Green (2005), for example, combined both approaches to estimate and explain score gains on IELTS writing tasks.

The findings of this line of research indicate that IELTS scores do change after instruction, but the direction and magnitude of score changes vary depending on language skill and learner characteristics (Green, 2005). Learners with lower scores before instruction tend to exhibit larger score gains than do those with initial higher scores. Some language skills (e.g., listening) showed greater score gains than others (e.g., writing) over the same period of instruction. This line of research provides important empirical evidence that supports the test's validity argument, namely that changes in test scores are associated with changes in L2 ability. However, as Green (2005) noted, these studies suggest also that individual score changes, whether gains or losses, might be due to factors other than changes in L2 ability, such as practice effects.

One limitation of previous studies on repeaters writing performance is that they looked only at changes in test scores and did not examine whether these score changes are associated with changes in the linguistic characteristics of candidates' texts. Additionally, these studies collected data at two time points in the form of pre- and post-tests (e.g., Elder and O'Loughlin, 2003) or on two testing occasions (e.g., Green, 2005). However, questions about the patterns of change in test performance and individual differences in change patterns over time can be answered only when at least three repeated measures of the same variable are available for each participant (Ross, 2005; Singer and Willett, 2003).

The current study aims to address these limitations by examining the linguistic characteristics of texts written in response to IELTS Writing Task 2 by candidates who took IELTS Academic three times.

1.2 Research on writing features distinguishing L2 proficiency levels

One approach to explain the meaning of L2 writing test scores is to examine the relationships between test scores and the linguistic and discourse characteristics of candidates' responses to writing tasks (e.g., Banerjee et al., 2007; Barkaoui, 2007, 2010b; Barkaoui and Knouzi, 2012; Cumming et al., 2005; Frase et al., 1999; Kennedy and Thorp, 2007; Mayor et al., 2007; Riazi and Knox, 2013). This approach is based on the assumption that the quality of test performance (as reflected in test scores) can be partially explained by examining the characteristics of the performance itself (Chapelle, 2008; Cumming et al., 2005; Taylor, 2004). Cumming et al. (2005), for example, compared the linguistic and discourse characteristics of scripts at different proficiency levels and on integrated and independent writing tasks in the New Generation TOEFL. They found that, regardless of task type, high-scoring scripts tended to be longer, demonstrate greater grammatical accuracy, and include a wider range of words, longer and more clauses, better quality claims, and more coherent summaries of source evidence, than did low-scoring scripts.

Three studies have recently examined the linguistic and discourse characteristics of IELTS Academic Writing Task 2 scripts written by candidates from different first-language (L1) backgrounds and assessed at different band levels. Mayor et al. (2007) examined the errors, complexity and discourse of Writing Task 2 scripts written by high-scoring (bands 7 and 8) and low-scoring (band 5) Chinese and Greek L1 candidates. They found that several features, including text length, formal error rate, sentence complexity, the use of the impersonal pronoun "one", thematic structure, argument genre and interpersonal tenor, were significant predictors of Writing Task 2 scores.

Banerjee et al. (2007) compared the linguistic characteristics of scripts written by Chinese and Spanish L1 candidates in response to IELTS Academic writing tasks 1 and 2 and scored at bands 3 to 8. Banerjee et al. examined several linguistic features, including cohesive devices, lexical variation and sophistication, syntactic complexity, and grammatical accuracy. They found that: (a) scripts at increasing ILETS band levels displayed greater lexical variation and sophistication; (b) gains in vocabulary are salient at lower levels, but other criteria become increasingly salient at higher levels; and (c) grammatical accuracy was a good discriminator of proficiency level regardless of task type and test taker L1.

More recently, Riazi and Knox (2013) compared the linguistic and discourse characteristics of IELTS Academic Writing Task 2 scripts written by three L1 candidate groups (European, Hindi and Arabic) assessed at three different band levels (5, 6 and 7). They found that scripts with higher band scores (6 and 7) tended to be longer and to include a higher proportion of low-frequency words, greater lexical diversity, and more syntactic complexity than did low-scoring scripts. However, high-scoring scripts were not necessarily more cohesive than low-scoring scripts.

The three studies also found significant differences in terms of some linguistic characteristics (e.g., lexical diversity) across L1 groups.

While the studies above have provided important insights concerning the nature and development of L2 proficiency and the effects of candidate and task factors on the characteristics of L2 writers' texts, they all adopted a cross-sectional approach, where writing samples by different candidates at different levels of L2 proficiency at one time point are analysed and compared in terms of their writing features. A longitudinal approach that focuses on *intra*-individual differences in test performance over time could contribute significantly to this line of research. Examining the scripts of candidates who take a L2 writing test more than once could help address questions concerning: (a) the nature and extent of differences and changes in the characteristics (e.g., linguistic accuracy, vocabulary use) of the scripts of test repeaters; and (b) the extent to which these differences and changes in script features are reflected in differences and changes in their writing scores. Here 'difference' refers to variation across candidates at one point in time, while 'change' refers to variation within the same candidate across time.

A challenge that faces studies on candidates' text features is to find the ideal group of measures that, when applied together, can detect variability in writing performance across individuals and time (Banerjee et al., 2007). To address this challenge, the current study adopts a detailed text analysis framework that builds on models of L2 ability, findings from previous research, and criteria on the IELTS rating scale for Writing Task 2 (see below).

2 THE PRESENT STUDY

This study aimed to examine the *patterns of changes over time* in the linguistic and discourse characteristics of texts written by IELTS repeaters in response to Writing Task 2.

Data consisted of the Writing Task 2 scores and scripts of three groups of candidates (*N*= 78) who took IELTS Academic three times (test occasions 1, 2 and 3). Candidate group was defined in terms of candidate Writing Task 2 score at test occasion 1 (i.e., band score 4, 5 or 6).

IELTS Writing Task 2 requires the candidate to write an argumentative text (in 40 minutes) that is at least 250-word long and in which the candidate presents a solution to a problem; presents and justifies an opinion; compares and contrasts evidence, opinions and implications; or evaluates and challenges ideas, evidence or an argument. The task assesses the candidate's ability to write a clear, relevant, well-organised argument, giving evidence or examples to support his/her ideas, and use English accurately.

Research questions

The study addressed the following research questions:

1. To what extent and how do the scripts of the three groups of candidates at test occasion 1 differ in terms of their linguistic characteristics?
2. To what extent and how do the linguistic characteristics of the repeaters' scripts change across test occasions?
3. To what extent and how does test repeaters' initial L2 writing ability (i.e., initial writing score) relate to changes in the linguistic characteristics of their scripts across test occasions?
4. To what extent and how do the linguistic characteristics of the repeaters' scripts relate to their writing scores across test occasions?

2.1 Sample and dataset

Data for the study were obtained from IELTS and consisted of individual biographical data (age, gender, L1 and country) and the IELTS Writing Task 2 scores and scripts for a purposive sample of 78 candidates who each took IELTS Academic three times.

The sample of candidates was selected based on their scores on IELTS Writing Task 2 at test occasion 1 (i.e., the first time they took the test). Specifically, three groups of candidates (*n*= 26 per group) were selected:

- group 1 included candidates whose scripts received a score of 4 at test occasion 1
- group 2 received a score of 5
- group 3 received a score of 6.

The sample consisted of 35 females (45%) and 43 males who came from 27 different countries, with the majority being from China (*n*=12), India (*n*= 12), Saudi Arabia (*n*= 9) and South Korea (*n*= 8).

They ranged between 16 and 52 years in terms of age (*M*= 25.65, *SD*= 6.63).

They spoke 23 different first languages, with the majority being L1 speakers of Arabic (*n*= 16), Chinese (*n*= 14), Korean (*n*= 8) and Punjabi (*n*= 7).

The study included 234 scripts (i.e., 26 candidates x 3 groups x 3 test occasions). Table 1 displays the sampling plan for the study. All participants took all three tests in 2013, but the length of period between the first and third test ranged between 14 and 219 days (i.e., 2 weeks to 7 months).

Table 2 displays descriptive statistics concerning the interval (in days) between test occasions. All scripts were handwritten by the candidates and then each script was typed (by IELTS staff) into a Word document, retaining the original script layout and mistakes.

Table 3 displays descriptive statistics for the overall and Writing Task 2 scores by candidate group and test occasion. It shows that the mean overall and writing scores for all three groups increased across test occasions. The inter-correlations (Pearson *r*) among writing task 2 scores across test occasions were high; they were *r*=.96 for occasions 1 and 2, .94 for occasions 2 and 3, and .90 for occasions 1 and 3.

Candidate group	Occasion 1	Occasion 2	Occasion 3	Total
4	26	26	26	78
5	26	26	26	78
6	26	26	26	78
Total	78	78	78	**234**

Table 1: Sample of scripts included in the study

	Test 1 to Test 2	Test 2 to Test 3	Test 1 to Test 3
M	57.29	53.64	110.94
SD	35.44	36.10	52.23
Min	7	5	14
Max	154	161	219

Table 2: Descriptive statistics for interval (in days) between test occasions

		Occasion 1		Occasion 2		Occasion 3	
Group		**Overall**	**Task 2**	**Overall**	**Task 2**	**Overall**	**Task 2**
4	M	4.73	4.00	4.85	4.63	5.25	5.33
	SD	.49	.00	.61	.27	.60	.45
5	M	5.56	5.00	5.81	5.56	6.12	6.25
	SD	.52	.00	.49	.22	.55	.35
6	M	6.79	6.00	7.04	6.62	7.27	7.19
	SD	.57	.00	.55	.26	.45	.35

Table 3: Descriptive statistics for Overall and Writing Task 2 scores by occasion and group

2.2 Data analyses

To examine the writing features of repeaters' Writing Task 2 scripts, the study used a detailed text analysis framework that builds on theory, previous research and criteria on the IELTS rating scale for Writing Task 2. Theoretically, the analytic framework is based on Connor and Mbaye's (2002) Model of Writing Competence. This model is based on Canale and Swain's (1980; Canale, 1983) model of Communicative Language Competence and includes: grammatical competence (e.g., grammar, lexis), discourse competence (e.g., coherence), sociolinguistic competence (e.g., register), and strategic competence (e.g., metadiscourse use). Connor and Mbaye argued that all four competencies should be reflected in any linguistic analysis of L2 learners' texts.

Table 4 presents the components of the Connor-Mbaye (2002) model (column 1), the main rating criteria for IELTS Writing Task 2 that correspond to each component (column 2), the specific writing features used in this study to operationalise each component (columns 3 and 4), and the computer programs used to estimate them (column 5).

The rating criteria for IELTS Writing Task 2 include: task response, coherence and cohesion, lexical resource, and grammatical range and accuracy (IELTS, 2009). The task response criterion is not included because none of the measures in Table 4 addresses this criterion (cf. Riazi and Knox, 2013). Like Riazi and Knox (2013), this study does not aim to examine linguistic features that perfectly match the IELTS Writing Task 2 rating criteria, but to examine variability in the linguistic and discourse characteristics of Writing Task 2 scripts across candidate groups and time.

Five computer programs were used to analyse the scripts in this study:

1. *Coh-Metrix* (Crossley et al., 2011; Graesser et al., 2004; McNamara et al., 2010)
2. *Criterion* (http://www.ets.org/criterion; Lim and Kahng, 2012; Ramineni et al., 2012; Weigle, 2010, 2011)
3. *L2 Syntactic Complexity Analyzer* (Lu, 2009, 2010, 2011)
4. *Multidimensional Analysis Tagger* (*MAT*; Nini, 2014)
5. *AntConc* **(**Anthony, 2012, 2013; Anthony and Bowen, 2013).

Coh-Metrix is web-based software that provides more than 100 computational linguistic indices of text coherence and cohesion, word diversity and characteristics, and syntactic complexity, measures that are considered to influence text quality. *Coh-Metrix* has been used in numerous studies to analyse texts written by L1 and L2 writers (e.g., Crossley and McNamara, 2011, 2014; Crossley et al., 2009, 2010, 2011; McNamara et al., 2010; Riazi and Knox, 2013).

The web-based program *Criterion* uses the e-rater scoring engine, the automated essay scoring system developed by Educational Testing Service (ETS), to examine text structure and linguistic accuracy (Ramineni et al., 2012; Weigle, 2010, 2011).

The *L2 Syntactic Complexity Analyzer* is a web-based program for identifying specific linguistic structures (e.g., sentences, clauses, T-units) in written texts (Lu, 2009, 2010, 2011).

Finally, *MAT* replicates Biber's (1988) tagger for the multidimensional functional analysis of English texts (Nini, 2014), while the concordance software *AntConc* allows the identification and counting of specific lexical items such as metadiscourse markers. The following paragraphs provide a detailed description and justification of each of the measures in Table 4.

2.2.1 Script linguistic characteristics

2.2.1.1 Grammatical

Fluency: Fluency refers to amount of production and is operationalised as the number of words per script. Several previous studies found that text length is one of the strongest predictors of L2 writing test scores (e.g., Cumming et al., 2005; Frase et al., 1999; Grant and Ginther, 2000; Mayor et al., 2007; Riazi and Knox, 2013).

Linguistic accuracy: Almost all studies that have examined the characteristics of L2 learners' texts examined accuracy, measured as the number of linguistic errors in a text (e.g., Cumming et al., 2005; Polio, 1997; Wolfe-Quintero et al., 1998). The web-based program, *Criterion* was used to identify, categorise and count the linguistic mistakes in each script. *Criterion* identifies four types of mistakes: grammar (e.g., sentence structure errors, pronoun errors, ill-formed verbs), usage (e.g., article errors, incorrect word forms), mechanics (e.g., spelling, punctuation), and style (e.g., passive voice, too many long sentences). An error ratio per 100 words (i.e., [total number of errors/total number of words] x 100) was computed for all errors and for each error type (i.e., grammar, usage, mechanics, and style) for each script.

Syntactic complexity: Syntactic complexity refers to the extent to which writers are able to incorporate increasingly large amounts of information into increasingly short grammatical units (Bardovi-Harlig, 1992; Polio, 2001). The developers of *Coh-Metrix* (e.g., Crossley, Greenfield and McNamara, 2008) noted that complex sentences are structurally dense or have many embedded constituents. *Coh-Metrix* was used to compute three indicators of syntactic complexity for each script: (a) left embeddedness, i.e., the mean number of words before the main verb of main clauses;
(b) noun-phrase (NP) density, which consists of the mean number of modifiers (e.g., determines, adjectives) per NP; and (c) syntactic similarity, which measures the uniformity and consistency of the syntactic constructions in the text.

Competence	IELTS rating criteria	Writing feature	Specific measure	Computer program
Grammatical	Grammatical range and accuracy	Fluency	Number of words per script	Coh-Metrix
		Accuracy	Number and distribution of four types of errors: grammar, usage, mechanics, and style.	Criterion
		Syntactic complexity	Left embeddedness; NP density; and syntactic similarity	Coh-Metrix
	Lexical resource	Lexical features	Lexical density Lexical variation Lexical sophistication	Coh-Metrix
Discourse	Coherence and cohesion	Cohesion and coherence	Connectives density Coreference cohesion Conceptual cohesion	Coh-Metrix
		Discourse structure	Organisation: Presence of 5 discourse elements (introductory material, thesis statement, main idea, supporting ideas, and conclusion) Development: Relative length of each discourse element	Criterion
Sociolinguistic		Register	Contractions, Passivisation, and Nominalisation	Multidimensional Analysis Tagger (MAT)
Strategic		Metadiscourse	Interactional metadiscourse markers	AntConc

Table 4: List of measures of the linguistic characteristics of repeaters' scripts

Coh-Metrix provides several indices of syntactic similarity; only one of them, mean sentence syntactic similarity for all combinations across paragraphs, was used in this study. Sentences with complex syntactic compositions have a higher ratio of constituents per NP than do sentences with simple syntax (Graesser et al., 2004). Generally, high syntactic similarity indices indicate less complex syntax (Crossley, Greenfield, and McNamara, 2008; Crossley et al., 2011).

Lexical features: Three lexical features were examined: lexical density, lexical variation, and lexical sophistication. *Lexical density* concerns the ratio of lexical words (i.e., nouns, verbs, adjectives, and adverbs) to the total number of words per script (Engber, 1995; Laufer and Nation, 1995; Lu, 2012). It was computed using *Coh-Metrix* by dividing the number of lexical words by the total number of words per script. Function or grammatical words (e.g., articles, prepositions, and pronouns) were not included in this analysis.

Lexical variation (or diversity) is often measured using Type-Token Ratio (TTR). TTR is the ratio of the types (the number of different words used) to the tokens (the total number of words used) in a text (Engber, 1995; Laufer and Nation, 1995; Lu, 2012; Malvern and Richards, 2002; Read, 2005). A high TTR suggests that the text includes a large proportion of different words (types), whereas a low ratio indicates that the writer makes repeated use of a smaller number of types. TTRs, however, tend to be affected by text length, which makes them unsuitable measures when there is much variability in text length (Koizumi, 2012; Lu, 2012; Malvern and Richards, 2002; McCarthy and Jarvis, 2010). The Measure of Textual and Lexical Diversity (MTLD), computed using *Coh-Metrix,* addresses this limitation since MTLD values do not vary as a function of text length, thus, allowing for comparisons between texts of considerably different lengths (Koizumi, 2012; McCarthy and Jarvis, 2010).

Lexical sophistication concerns the proportion of relatively unusual, advanced, or low-frequency words to frequent words used in a text (Laufer and Nation, 1995; Meara and Bell, 2001). Two measures were used to assess lexical sophistication, average word length (AWL) and word frequency, both computed by *Coh-Metrix*. AWL is computed by dividing the total number of letters by the total number of words for each script (Biber, 1988; Cumming et al., 2005; Engber, 1995; Frase et al., 1999; Grant and Ginther, 2000). Higher AWL values indicate more sophisticated vocabulary use.

Word frequency, measured using the mean CELEX word frequency score for content words, refers to how often particular content words occur in the English language (Graesser et al., 2004). The CELEX frequency score is based on the database from the Centre of Lexical Information (CELEX) which consists of frequencies taken from the early 1991 version of the COBUILD corpus of 17.9 million words (see Crossley et al., 2007, 2008). Research suggests that advanced L2 learners are more likely to comprehend and use lower-frequency words than do learners with low L2 proficiency (Bell, 2003; Crossley et al., 2010; Ellis, 2002; Meara and Bell, 2001).

2.2.1.2 Discourse

To examine discourse, each script was computer-coded in terms of several coherence and cohesion features and various aspects of discourse structure.

Coherence and cohesion: Using *Coh-Metrix*, each script was computer-analysed in terms of connectives density, coreference cohesion, and conceptual cohesion. *Connectives* provide explicit cues to the types of relationships between ideas in a text, thus, providing important information about a text's cohesion, organisation, and quality (Crismore et al, 1993; Halliday and Hasan, 1976). *Coh-Metrix* provided an incidence score (occurrence per 1000 words) for all connectives (i.e., causal, additive, temporal and clarification connectives) for each script. *Coreference cohesion* occurs when a noun, pronoun, or noun phrase refers to another constituent in the text (Crossley et al., 2007, 2009, 2011; McNamara et al., 2010). *Coh-Metrix* provides indices concerning several types of coreferentiality. These indices, however, were highly inter-correlated ($r > .70$). Consequently, only one of them was included in the study: argument overlap for adjacent sentences, which measures how often two adjacent sentences share common arguments (i.e., nouns, pronouns, and noun phrases).

Conceptual cohesion concerns the extent to which the content of sentences or paragraphs is similar semantically or conceptually. The main measures of this variable are based on Latent Semantic Analysis (LSA). LSA is a statistical, corpus-based technique that provides an index of local and global conceptual cohesion and coherence between parts of a text by considering similarity in meaning, or conceptual relatedness, between and within parts of a text (i.e., sentences, paragraphs) (Crossley et al., 2008; Crossley et al., 2009, 2011; Foltz, Kintsch, and Landauer, 1998; Graesser et al., 2004; Landauer, Foltz, and Laham, 1998; McNamara, Cai, and Louwerse, 2007). Unlike lexical markers of coreferentiality (i.e., noun and argument overlap), LSA provides for the tracking of words that are semantically similar, but may not be related morphologically (Landauer and Dumais 1997; Landauer, Foltz and Laham 1998).

Text cohesion (and sometimes coherence) is assumed to increase as a function of higher conceptual similarity between text constituents (Crossley, Louwerse, et al., 2007; Landauer et al., 2007). LSA has been used in previous studies to estimate L1 text coherence and to grade L1 essays in English composition (e.g., Landauer et al., 2007). *Coh-Metrix* was used to compute two LSA scores for each script: (a) mean LSA overlap for adjacent sentences (i.e., how similar a sentence is to adjacent sentences) and (b) mean LSA overlap for adjacent paragraphs (i.e., how similar a paragraph is to adjacent paragraphs).

Discourse structure: To examine text structure, the web-based program *Criterion* was used to measure the organisation and development of each script. *Criterion* automatically identifies sentences in each script that correspond to each of five discourse elements: introductory material (background), thesis statement, main idea, supporting ideas, and conclusion (Ramineni et al., 2012; Weigle, 2011). For organisation, *Criterion* identifies whether each script includes each of the five discourse elements. Development is measured by computing the relative length of each discourse element (Ramineni et al., 2012; Weigle, 2011). This was done by dividing the number of words assigned to each discourse element by the total number of words in each script and multiplying the result by 100.

2.2.1.3 Sociolinguistic

Most studies on sociolinguistic competence in the context of L2 writing focus on register, particularly the use of written (or formal) and spoken (or informal) features (e.g., Biber, 1988; Chang and Swales, 1999; Grant and Ginther, 2000; Hinkel, 2003; Shaw and Liu, 1998). Some of these studies counted features that are associated with informal speech style (e.g., personal pronouns, direct questions, exclamations, simple syntax, contractions, broad references) (e.g., Chang and Swales, 1999; Hinkel, 2003), while other studies counted the number of formal features that indicate an academic style (e.g., passive voice, formal vocabulary, nominalisation, complex syntax, hedging, rich modification) (e.g., Grant and Ginther, 2000; Shaw and Liu, 1998). Some of these features are considered under other categories above (e.g., lexical features, syntactic complexity).

Additionally, this study examined three specific features in relation to register: *contractions* (e.g., *won't*), *passivisation* (i.e., number of *by*- and agentless passive constructions), and *nominalisation* (i.e., number of nouns ending in –ance/-ence, -cy, -ion, -ism, -ist, -ity, -ive, -ment, -ness, -ure). Grant and Ginther (2000) found that the frequency of passivisation and nominalisation differed significantly across TOEFL score levels. The computer program *Multidimensional Analysis Tagger* (*MAT*; Nini, 2014) was used to compute contraction, passivisation, and nominalisation ratios (per 100 words) for each script.

Marker	Function	Examples
Interactional	Involve the reader in the text	
Hedges	Indicate degree of confidence in a proposition; withhold writer's full commitment to proposition	Might; perhaps; possibly; from my perspective, generally speaking, in my view
Boosters	Indicate certainty. Emphasise force or writer's certainty in proposition	In fact; definitely, certainly, no doubt, for sure, really
Attitude markers	Express writer's attitude to proposition	Unfortunately; I agree; appropriate, disappointing, dramatic, I believe
Self mentions	Explicit reference to author(s)	I; we; my; me; our
Engagement markers	Explicitly address, refer to or build relationship with reader	Let's, you, your, you can see that

(Source: Hyland, 2005, p. 49)

Table 5: Interactional metadiscourse markers

2.2.1.4 Strategic

One feature was examined in relation to strategic competence: use of *interactional metadiscourse markers* (Hyland, 2005; Hyland and Tse, 2004; Intaraprawat and Steffensen, 1995). As Hyland (2005) explained, metadiscourse refers to "the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer to express a viewpoint and engage with readers as members of a particular community" (p. 37). Connor and Mbaye (2002) noted that metadiscourse markers used in writing are similar to "repair strategies in spoken discourse" (p. 267). Based on previous theoretical models and empirical research, Hyland (2005) developed a classification scheme of metadiscourse markers that distinguishes between interactive and interactional metadiscourse resources. Interactive metadiscourse markers enable the writer to organise and guide the reader through their texts. They include, for example, transition markers (e.g., conjunctions, comparisons), frame markers (e.g., sequencing, topic shifts), and code glosses (e.g., for example, that is). Interactive metadiscourse markers are closely related to coherence and cohesion and are addressed under discourse competence above (e.g., connectives density).

Interactional choices, on the other hand, "focus more directly on the participants of the interaction, with the writer adopting an acceptable persona and a tenor consistent with the norms of the community" (p. 53). Interactional resources allow the writer to involve readers and alert them to his/her perspective towards both propositional information and readers themselves.

As Table 5 shows, interactional devices include hedges, boosters, attitude markers, self-mentions and engagement markers. The concordance software *AntConc* **(**Anthony, 2012, 2013; Anthony and Bowen, 2013) was used to identify the interactional metadiscourse markers used in each script. Next, the density of metadiscourse markers for each script was computed by dividing the total number of different markers by the total number of T-units per script (following Intaraprawat and Steffensen, 1995). A T-unit is defined as "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" (Hunt, 1970, p. 4, cited in Lu, 2011, p. 44). The web-based program *L2 Syntactic Complexity Analyzer* (Lu, 2009, 2010, 2011) was used to estimate the number of T-units for each script.

As mentioned earlier, several computer programs were used to analyse the scripts in terms of the various linguistic and discourse features listed above. A major issue when using computer programs to analyse texts written by L2 learners is that these texts often include several linguistic and other inaccuracies (e.g., misspelled words, incorrect punctuation). For example, some computer programs may not accurately identify and estimate the frequency of linguistic and discourse features if a script includes several and/or severe spelling and/or punctuation mistakes.

To address this problem, previous studies corrected the spelling and punctuation mistakes of these texts before conducting computer analyses of L2 learners' texts. For example, Crossley, Salsbury, McNamara and Jarvis (2010) and Crossley and McNamara (2014) corrected all texts in terms of spelling, while Kormos (2011) corrected punctuation mistakes before analysing the L2 learners' texts in their studies using *Coh-Metrix* in order to avoid ambiguous words and structures (cf. Riazi and Knox, 2013). Similarly, Lu (personal communication, 5 June 2014) recommended "correcting (a) obvious spelling mistakes and (b) punctuation that affect sentence segmentation (e.g., run-on sentences, sentences not ending with appropriate punctuation)" before analysing L2 learners' texts with the *L2 Syntactic Complexity Analyzer*.

Consequently, it was decided to create a new version of each script in this study with corrected spelling and punctuation mistakes. Only spelling mistakes that were detected by the spell checker in *Microsoft Word* and whose meaning can be understood from the context were corrected. For example, the word 'phenomenon', which is misspelled in the phrase "this phirominon makes the more retired people..." (script 415A) was corrected, but 'Jagh' in the phrase "if the crimes stay in the Jagh..." (script 404B) was not corrected because its meaning was not clear from the context. Next, punctuation mistakes such as run-on sentences and missing final punctuation were fixed. Analyses were then conducted on the corrected scripts, except when examining linguistic accuracy (using the web-based tool *Criterion*). For *Criterion* analyses, the original scripts were used to identify and classify language mistakes.

In order to evaluate the impact of the revisions made to the original scripts in terms of spelling and punctuation on the various linguistic indices in Table 4 above, both the original and corrected scripts were submitted to the same computer analyses. The results were then compared across versions for each index using mixed-design ANOVAs, with three independent variables: script version, candidate group, and test occasion. Only one measure – left embeddedness – showed significant differences across versions. Specifically, the original scripts had a higher mean index of left embeddedness (M= 5.09) than did the corrected scripts (M= 4.61). There were no significant interaction effects between script versions, candidate group and test occasion on left embeddedness.

2.2.2 Statistical analyses

Data for this study consisted of the Writing Task 2 scores and the measures of linguistic and discourse features listed in Table 4 above for each script for each candidate at each test occasion. Several analyses were conducted to address the research questions of the study.

First, descriptive statistics (e.g., means, standard deviations) for each linguistic and discourse feature in Table 4 were computed for all candidates and across test occasions and candidate groups.

Second, to address research question 1 concerning differences between the linguistic characteristics of the scripts of the three candidate groups (i.e., band scores 4, 5, and 6 at test occasion 1), univariate analysis of variance (ANOVA) was conducted for each linguistic measure in Table 4 with candidate group as the independent variable and the linguistic index as the dependent variable. When ANOVA results were significant, follow-up pairwise comparisons using a Bonferroni correction were conducted to compare pairs of candidate groups. For the presence of discourse structure elements (i.e., introduction, thesis statement, etc.), Chi-square (X^2) tests were conducted for each discourse element in order to examine the association between candidate group and the presence or absence of each discourse structure element. For all ANOVA analyses, only statistics (i.e., *F, df*, effect size) for significant effects ($p < .05$) are reported. Statistics for non-significant effects are not reported. Furthermore, *partial eta-squared* (*partial* $\dot{\eta}^2$) is used as a measure of effect size. *Partial* $\dot{\eta}^2 \geq .01$ indicates a small effect size; *partial* $\dot{\eta}^2 \geq .09$ indicates a medium effect; and *partial* $\dot{\eta}^2 \geq .25$ indicates a large effect (Field, 2009).

Third, the autocorrelations (Pearson r) of each measure of a linguistic feature with itself across successive test occasions (e.g., the correlations of lexical density with itself for time 1 and time 2 and for time 2 and time 3) were computed to find out whether and to what extent the order of candidates relative to each other changed across test occasions for each linguistic feature in the study.

Fourth, to examine the differences and changes in the linguistic and discourse characteristics of the scripts within and across test occasions and their relationships to differences in candidate initial writing abilities, multilevel modelling (MLM), using the computer program *HLM6* (Raudenbush, Bryk, Cheong and Congdon, 2004), was employed.

MLM is a family of statistical models for analysing data with nested structure (Barkaoui, 2013, 2014; Hox, 2002; Luke, 2008). MLM views repeated-measures observations as nested within individual cases and distinguishes between two levels of analysis: level-1 observations (i.e., test occasion) nested in level-2 units (i.e., candidate). Given an outcome variable, such as a linguistic feature index (e.g., fluency), the level-1 equation examines whether and how the outcome changes within each candidate over time. The level-1 equation includes two main parameters of change for each linguistic feature for each candidate: initial status (i.e., the intercept of the candidate's trajectory) and the rate of change (i.e., the slope of the candidate's trajectory) over time.

Trends in change in a linguistic feature can be tested to find out if they are linear or non-linear and parallel change processes can be examined as time-varying predictors (Luke, 2008; Preacher et al., 2008; Ross, 2005). Time-varying (or *intra*-individual) predictors are variables whose value changes over time such as candidate age and L2 proficiency. They are included as level-1 predictors in MLM. In contrast, time invariant (or *inter*-individual) predictors are variables that are constant across time such as candidate L1 and gender. They are included as level-2 predictors in MLM (Luke, 2008). The change trajectory within individuals can vary across individuals in terms of initial status (intercept) and/or rate of change (slope) (Luke, 2008). At level 2, candidates' initial status (i.e., their intercepts) and change rates (i.e., their slopes) serve as dependent variables, and candidate factors (e.g., writing score at occasion 1, gender) or important covariates are entered as predictor variables. The level-2 models, thus, examine the factors influencing the rate and shape of change in the outcome (e.g., fluency) over time.

Following Hox (2002), several MLM models were developed and evaluated for each linguistic feature separately before estimating the final model for that feature.

First, a null model with no predictors was examined to estimate the proportion of variance between candidates versus variance across test occasions (i.e., within candidate) for each linguistic feature. Model 2 included occasion as a predictor at level 1 to estimate the amount of change over time in each linguistic feature (i.e., research question 2). The slope of the occasion variable was allowed to vary across candidates in order to estimate the extent to which both initial status (i.e., intercept) and rate of change over time (i.e., slope) for each linguistic feature varied across candidates. Model 3 added one level-2 predictor, candidate group, in order to estimate the relationships between candidate group and differences in each of the linguistic characteristics of their scripts at time 1 (i.e., research question 1). Finally, Model 4 included cross-level interactions with occasion. Specifically, candidate group was added to estimate the relationship between candidate initial writing ability and the rate of change in each of the linguistic features over time (i.e., research question 3). Given the small sample of candidates included in the study, Restricted Maximum Likelihood (RML) was used for estimating all parameters as recommended by Hox (2002) and Luke (2004).

In all analyses, occasion was uncentered, with occasion 1 coded 0 so the intercept can be interpreted as the expected (average) outcome at occasion 1. Writing Task 2 score at time 1 was also uncentered (with band 4 coded 0, band 5 =1, and band 6=2). For each model, two main indices were examined: the deviance statistic, which compares the fit of multiple models to the same dataset, and significance tests for individual coefficients (Hox, 2002; Luke, 2004).

Based on the results of these different models, a final model was built for each linguistic feature. Section 3.2.1 illustrates all the steps and decisions involved in building and evaluating MLM models for fluency. However, to keep the report short, only the results for three MLM models (Model 1, Model 2 and final model) are discussed for the other linguistic features in the findings section. In all cases, the final model is compared to Model 1 in terms of fit statistics.

Finally, to address research question 4, concerning the relationships between the linguistic and discourse characteristics of candidates' scripts, on the one hand, and their Writing Task 2 scores, on the other, correlational analyses and MLM (using *HLM6*) were employed. Pearson *r* correlations between candidate Writing Task 2 scores and each linguistic feature were computed for each test occasion. To assess whether the strength of the association between a given linguistic feature and writing scores varied significantly across test occasions, the interactive calculator developed by Lee and Preacher (2013) to test the equality of two correlation coefficients obtained from the same sample was used. Additionally, correlations (Pearson *r*) among measures that assess the same linguistic feature were examined for each test occasion in order to identify features that are highly correlated (i.e., $r \geq .70$). If two or more measures were highly correlated, only one of them was retained for MLM analyses since highly correlated measures are likely to tap the same construct. Including only one index among measures that are highly inter-correlated also reduces the threat of multicollinearity.

Next, several MLM models were built and evaluated to assess the relationships between the linguistic and discourse features of the scripts and writing scores over time. RML was used in all analyses. First, a null model with no predictors was examined to estimate the proportion of variance between candidates versus variance across test occasions (i.e., within candidate) in writing scores. A second model included occasion as a predictor at level 1 in order to estimate the amount of change in writing scores over time. The occasion slope was allowed to vary across individuals to estimate the extent to which the rate of change in writing scores over time (i.e., slope) varied across candidates. Occasion was uncentered, with time 1 coded 0 so the intercept can be interpreted as the expected (average) writing score at time 1. Model 3 examined whether change in each of the linguistic features is significantly associated with change in writing scores over time. To reduce the number of linguistic features, only those features that were found to be significantly correlated with writing scores (based on correlational analyses) were included in MLM analyses. In order to make the interpretation of the intercept easier, all the linguistic measures were grand-mean cantered (i.e., the variable mean across all test takers and occasions).

Next, several sub-models were specified and tested to examine whether the relationships between each linguistic feature and writing scores varied significantly across candidates. In each of these sub-models, the relationship between one linguistic feature and writing scores was allowed to vary across candidates. Two statistics were examined to assess whether the association varied significantly across candidates: model fit indices (i.e., deviance statistics) and chi-square (X^2) tests which test whether a coefficient has a significant random variance across level-2 units (Barkaoui, 2013; Hox, 2002; Luke, 2004). Based on the results of these different models, only linguistic features that have significant association with writing scores were retained in the final model.

Because of the small number of cases included in the study, neither practice effects (i.e., number of previous tests taken and length of interval between test occasions) nor candidate variables (i.e., gender, age, L1) were included in the MLM analyses above. Nevertheless, the MLM models specified above allow for the examination of the extent to which changes in the linguistic features and scores of repeaters' scripts over time as well as the relationships between script characteristics and scores varied significantly across candidates. Future studies with larger samples could examine the effects of candidate and other factors on changes and differences in the characteristics and scores of repeaters' scripts.

3 FINDINGS

This section reports the results of the various analyses described above. The first subsection reports findings from ANOVA analyses concerning research question 1. Given that the MLM analyses for research questions 2 and 3 were conducted for each linguistic feature separately, the results for these questions are organised and reported by linguistic feature. That is, research questions 2 and 3 will be addressed separately for each linguistic feature. Next, findings concerning research question 4 are presented. Section 4 summarises the key findings across all linguistic features in relation to each research question.

3.1 Differences in the linguistic characteristics of scripts at different band levels at test occasion 1

Table 6 displays descriptive statistics for all linguistic measures across candidate groups (defined in terms of candidate Writing Task 2 score at time 1) for test occasion 1. Fluency was measured in terms of the total number of words per script. ANOVA detected significant differences across groups (*F*[2, 75]= 9.46, *p*<.05, $\acute{\eta}^2$= .20) on fluency. Follow-up pairwise comparisons (with Bonferroni correction) indicated that there were significant differences between candidates scoring 4 at time 1, on the one hand, and those scoring 5 and 6, on the other. As Table 6 shows, candidates who scored 4 at time 1 wrote on average significantly shorter texts (*M*= 228.92 words) than did those scoring 5 (*M*= 255.81) and 6 (*M*= 291.73) at time 1. The difference between candidates scoring 5 and those scoring 6 at time 1 was not significant.

Accuracy was measured (using *Criterion*) in terms of the frequencies of four types of errors (grammar, usage, mechanics and style) per 100 words. A ratio of total errors per 100 words was also computed for each script. ANOVA detected significant and large main effects for candidate group (F[2, 75]= 17.24, *p*<.05, $\acute{\eta}^2$= .31) for the ratio of all errors. Follow-up pairwise comparisons indicated that candidates scoring 4 at time 1 made significantly more errors (*M*= 21 errors per 100 words) than did those scoring 5 (*M*= 14) and 6 (*M*= 11) at time 1; the difference between the latter two groups was not significant. As Table 6 shows, for all error types (i.e., grammar, usage, mechanics, and style), candidates scoring 4 at time 1, on average, made more errors per 100 words than did those scoring 5 and 6.

Three measures of syntactic complexity were examined: left embeddedness (i.e., the mean number of words before the main verb of main clauses), NP density (i.e., mean number of modifiers per NP), and mean sentence syntactic similarity for all combinations across paragraphs. Table 6 displays descriptive statistics for each of the three measures across candidate groups at test occasion 1. ANOVA detected no significant main effects for candidate group on the complexity measures except for a small effect on NP density (F[2, 75]= 3.06, *p*=.05, $\acute{\eta}^2$= .08). Follow-up pairwise comparisons indicated that there was a significant difference in terms of NP density between candidates scoring 6 at time 1 (*M*= .83) and those scoring 4 (*M*= .72). There were no significant differences between scripts scoring 5 and 6 at test occasion 1 (*p*>.05).

Three lexical features were examined: lexical density (i.e., ratio of lexical words to total number of words per script), lexical variation (Measure of Textual and Lexical Diversity, MTLD), and lexical sophistication (average word length [AWL] and word frequency). ANOVAs indicated that there were significant main effects for candidate group on each of the four lexical measures at test occasion 1: lexical density(F[2, 75]= 5.58, *p*<.05, $\acute{\eta}^2$= .13), lexical variation (MTLD) (F[2, 75]= 10.09, *p*<.05, $\acute{\eta}^2$= .21), AWL (F[2, 75]= 29.48, *p*<.05, $\acute{\eta}^2$= .44), and word frequency (F[2, 75]= 18.44, *p*<.05, $\acute{\eta}^2$= .33).

Follow-up pairwise comparisons indicated that for each of the four lexical measures, there were significant differences between candidates scoring 6 at time 1, on the one hand, and those scoring 4 and 5, on the other. The difference between the latter two groups was not significant for any of the lexical measures. As Table 6 shows, on average, candidates scoring 6 at time 1 had higher lexical density indices (*M*= .54) than did those scoring 4 (*M*= .52) and 5 (*M*= .51). Similarly for lexical variation, candidates scoring 6 had significantly higher MTLD indices (*M*= 86.68), than did those scoring 4 (*M*= 64.26) and 5 (*M*= 69.98). Furthermore, candidates scoring 6 at time 1, generally, used longer words (*M*= 5.06 letters per word) and less frequent words (*M*= 2.33) than did candidates scoring 4 and 5 as shown in Table 6. Overall, candidates scoring 6 at time 1, on average, used significantly more content words, more diverse vocabulary, longer words, and more low-frequency words than did candidates with lower writing scores (4 and 5) at time 1.

Coherence and cohesion were measured in terms of three features: connectives density (i.e., number of connectives per 1000 words), coreference cohesion (i.e., argument overlap for adjacent sentences), and conceptual cohesions (i.e., mean LSA overlap for adjacent sentences and mean LSA overlap for adjacent paragraphs). ANOVA detected a significant but small effect for candidate group on mean LSAP overlap for adjacent paragraphs only (F[2, 75]= 3.06, *p*=.05, $\acute{\eta}^2$= .08). Follow-up pairwise comparisons indicated that there was a significant difference between candidates scoring 6 at time 1 (*M*= .41) and those scoring 4 (*M*= .32). There were no significant differences between scripts scoring 5 and 6 at test occasion 1 (*p*>.05).

Candidate group	**4** (n= 26)		**5** (n= 26)		**6** (n= 26)		**ANOVA**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**F**	**p**
Fluency								
Words per script	228.92	59.07	274.46	66.56	304.31	62.93	9.46	.00
Accuracy (per 100 words)								
All errors	21	8	14	6	11	6	17.24	.00
Grammar	2	1	1	1	1	1	8.64	.00
Usage	3	2	2	2	2	1	3.29	.04
Mechanics	6	4	3	2	3	2	12.12	.00
Style	10	7	7	5	5	5	5.72	.00
Syntactic complexity								
Left embeddedness	4.74	2.62	4.19	1.14	4.83	1.57	.88	.42
Syntactic similarity	0.10	0.03	0.11	0.03	0.10	0.02	3.06	.05
NP density	0.72	0.12	0.77	0.16	0.83	0.17	.11	.89
Lexical features								
Lexical density	.52	.05	.51	.04	.54	.03	5.58	.01
MTLD	64.26	16.21	69.98	19.06	86.68	20.57	10.09	.00
AWL	4.49	0.28	4.63	0.31	5.06	0.24	29.48	.00
Word frequency	2.53	0.13	2.49	0.13	2.33	0.11	18.44	.00
Coherence and cohesion								
All connectives	102.81	22.1	105.39	20.70	103.49	17.68	.11	.89
Argument overlap	0.55	0.23	0.55	0.17	0.55	0.16	.00	1.00
LSA overlap, sentences	0.20	0.09	0.18	0.07	0.19	0.05	.21	.81
LSA overlap, paragraphs	0.32	0.16	0.40	0.16	0.41	0.09	3.06	.05
Development (percentage)								
Introduction	3.57	5.26	7.65	7.99	8.00	7.36	3.25	.04
Thesis	11.16	8.82	11.57	7.46	8.27	7.64	1.31	.27
Main idea	20.33	14.41	17.18	12.59	18.33	9.09	.44	.64
Supporting ideas	51.18	19.29	48.21	14.96	50.56	12.30	.26	.77
Conclusion	9.07	7.55	13.57	10.60	12.73	6.01	2.18	.12
Register (per 100 words)								
Contractions	0.58	0.67	0.47	0.84	0.06	0.20	4.90	.01
Passivisation	0.40	0.63	0.42	0.45	1.22	0.83	13.40	.00
Nominalisation	2.12	1.31	4.36	3.30	4.18	2.45	6.47	.00
Metadiscourse markers								
Interactional	1.17	0.62	1.12	0.69	1.07	0.50	.16	.85
Hedges	0.28	0.25	0.22	0.15	0.40	0.26	4.63	.01
Boosters	0.08	0.07	0.11	0.11	0.13	0.13	1.33	.27
Attitude markers	0.22	0.2	0.18	0.12	0.21	0.13	.62	.54
Self mention	0.38	0.27	0.40	0.35	0.20	0.21	3.76	.03
Engagement markers	0.21	0.24	0.23	0.30	0.13	0.20	1.06	.35

Table 6: Descriptive statistics for linguistic features by candidate group at test occasion 1

Candidate group	**4** (n= 26)		**5** (n= 26)		**6** (n= 26)		**Chi-square test**	
	f	**%**	**f**	**%**	**f**	**%**	**X^2**	**p**
Introduction	10	38	15	58	19	73	6.63	.04
Thesis	21	81	23	88	18	69	2.98	.22
Main idea	25	96	25	96	26	100	1.03	.60
Support	26	100	26	100	26	100	NA	NA
Conclusion	18	69	22	85	24	92	4.88	.09

Table 7: Descriptive statistics for organisation by candidate group at test occasion 1

As noted above, *Criterion* was used to examine script organisation, i.e., whether each script included each of five discourse elements (introductory material, thesis statement, main idea, supporting ideas, and conclusion), and development (i.e., the percentage of the script assigned to each discourse element included in the script). Table 7 displays descriptive statistics for organisation by candidate group at test occasion 1. All candidates included supporting ideas in their scripts, but the percentage of candidates who included other discourse elements varied across candidate groups. In order to examine whether there are significant associations between candidate group and the presence or absence of each discourse element, chi-square (X^2) tests were conducted for each discourse element separately.

The results (Table 7) indicated that there was a significant association of candidate group with introduction (X^2= 6.63, *df*.= 2 *p*<.05), but not for the other discourse elements. As Table 7 shows, a significantly higher proportion of candidates scoring 6 at time 1 included an introduction (73%) than did candidates scoring 5 (58%) and 4 (38%). The same pattern was true for the conclusion as well, with more scripts scoring 6 (92%) including a conclusion compared to those scoring 5 (85%) and 4 (69%), but this association was not statistically significant (*p*>.05).

As for development (i.e., length of each discourse element), ANOVA detected a significant but small effect for candidate group only on the length of the introduction (F[2, 75]= 3.25, *p*=.05, η^2= .08). Specifically, as Table 6 shows, candidates scoring 5 and 6 devoted significantly a higher proportion of their texts (about 8%) to the introduction compared to those scoring 4 (*M*= 3.57%).

Three features were examined in relation to register, the ratios of contractions, passive constructions, and nominalisations per 100 words. Contractions are associated with informal speech, while passive constructions and nominalisations are associated with formal academic style. Table 6 reports descriptive statistics for each feature across candidate groups at test occasion 1.

ANOVAs indicated that there was a significant medium to large effect for candidate group on contractions (F[2, 75]= 4.90, *p*<.05, η^2= .12), passive constructions (F[2, 75]= 13.40, *p*<.05, η^2= .26), and nominalisation (F[2, 75]= 6.47, *p*<.05, η^2= .15). Follow-up pairwise comparisons indicated that candidates scoring 6 at time 1 used significantly fewer contractions (*M*= .06 contractions per 100 words) and more passive constructions (*M*= 1.22 per 100 words) than did those scoring 5 (*M*= .47 and *M*= .42) and 4 (*M*= .58 and *M*= .40, respectively). For nominalisation, as Table 6 shows, candidates scoring 4 at time 1 used significantly fewer nominalisations (*M*= 2.12 per 100 words) than did those scoring 5 (*M*= 4.36) and 6 (*M*= 4.18).

Finally, one feature was examined under strategic competence: use of interactional metadiscourse markers. Table 6 displays descriptive statistics for interactional metadiscourse markers, as well as their subcategories, across candidate groups for test occasion 1. ANOVA detected no significant effects for candidate group on the use of interactional metadiscourse markers. However, there were significant medium effects for candidate group on the use of hedges (F[2, 75]= 4.63, *p*<.05, η^2= .11) and self-mentions (F[2, 75]= 3.76, *p*<.05, η^2= .09). Specifically, candidates scoring 6 at time 1 used significantly more hedges (*M*= .40 hedges per T-unit) and fewer self-mentions (*M*= .20 hedges per T-unit) than did those scoring 5 (*M*= .22 and *M*=40, respectively). The differences between candidates scoring 4 and the other two candidate groups were not significant for any of the metadiscourse measures.

3.2 Changes in the linguistic characteristics of repeaters' scripts across test occasions

This section reports the MLM results concerning research questions 1 to 3. As noted above, the findings are organised and reported by linguistic feature.

3.2.1 Fluency

Table 8 displays descriptive statistics for fluency across test occasions and candidate groups (defined in terms of candidate Writing Task 2 score at time 1). It shows that candidates who scored 4 at time 1 wrote on average shorter texts than did those scoring 5 and 6 at each of the test occasions. Note also that the scripts at each test occasion are longer than scripts at the previous occasion. Thus, the scripts at test occasion 3 were, on average, longer (*M*= 312.51 words) than those produced at test occasion 2 (*M*= 289.86 words), which were in turn longer than those produced at test occasion 1 (*M*= 269.23 words).

However, the differences between candidate groups in terms of script length decreased over time. For example, the difference in terms of the number of words per script between candidates scoring 4 and those scoring 6 decreased from an average of 62.81 words at test occasion 1, to 39.35 words at test occasion 2, to 27.69 words at test occasion 3.

Finally, the autocorrelations (Pearson *r*) of fluency over time were positive and significant (*r* =.62 for occasions 1 and 2 and .62 for occasions 2 and 3; both *p*<.01). Generally, candidates who produced longer scripts at each test occasion produced longer scripts at the following test occasion and vice versa.

Table 9 displays the results for the various MLM models that were examined for fluency. The table includes three sets of statistics: fixed effects, random effects, and model fit. *Fixed effects* can be interpreted in the same way as coefficients in multiple regression analysis. They include the intercept of the outcome and a slope for each predictor; the slope indicates the strength of the association between each predictor and the outcome (controlling for the effects of other predictors in the model). MLM uses *t*-tests to test whether a fixed effect (i.e., intercept or slope) significantly departs from zero. As a rule of thumb, a coefficient reaches significance at *p* < .05 when its estimate is twice as large as its standard error (SE) (Hox, 2002).

Random effects refer to the magnitude of variance in coefficients (i.e., intercept or slope) across candidates. Chi-square (X^2) tests are used to test whether a random effect significantly departs from zero. A significant random effect indicates that the coefficient (intercept or slope) varies significantly across candidates.

Candidate Group	**4**		**5**		**6**		**Total**	
Occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1	228.92	59.07	274.46	66.56	304.31	62.93	269.23	69.50
Occasion 2	255.81	45.79	296.92	55.74	316.85	59.13	289.86	58.98
Occasion 3	291.73	65.58	313.81	56.65	332.00	69.01	312.51	65.26
Total	**258.82**	**62.29**	**295.06**	**61.26**	**317.72**	**64.01**	**290.53**	**66.84**

Table 8: Descriptive statistics for fluency by candidate group and test occasion

	Model 1	**Model 2**	**Model 3**	**Model 4**
Fixed effects (SE)				
Intercept	290.53** (6.15)	268.89** (7.41)	238.94** (8.81)	230.67** (10.07)
Candidate group			29.95** (6.75)	38.23** (7.69)
Occasion		21.64** (3.77)	21.64** (3.77)	30.42** (6.65)
Candidate group				-8.78 (4.63)
Random effects				
Between-candidate	2235.96	3214.04	2305.16	2271.22
X^2 (df.)	306.51** (77)	297.71** (77)	233.13** (76)	229.94** (76)
Occasion Slope		450.97	451.12	413.03
X^2 (df)		128.61** (77)	128.62** (77)	122.66** (76)
Within-candidate	2250.47	1345.58	1345.49	1345.58
Model fit				
Deviance (parameters)	2569.72 (2)	2525.54 (4)	2501.39 (4)	2494.71 (4)
Model comparison: X^2 (df.)		44.18** (2)	68.33** (2)	75.01** (2)

* *p*<.05; ** *p*<.01; N=78

Table 9: MLM results for fluency

Finally, *model fit* is assessed using the deviance statistic. The deviance for any one model cannot be interpreted directly, but it can be used to compare the overall fit of multiple models to the same dataset (Barkaoui, 2013; Hox, 2002; Luke, 2004). Generally, models with a lower deviance fit better than models with a higher deviance (Hox, 2002). The chi-square (X^2) difference test was used to assess whether more complex models (i.e., models including more parameters) improve model fit significantly compared to less complex ones.

As noted above, Model 1 assessed the proportion of variance between candidates versus variance across test occasions (i.e., within candidate) in fluency. The results for Model 1 indicated that there was approximately the same intra-individual variability (2250.47) as inter-individual variability (2235.96) in fluency. The total fluency measure variance is (2250.47+2235.96=) 4486.43. The interclass correlation (ICC), or the proportion of variance at the person level, is estimated as (2235.96/4486.43=) .50.

In other words, half (50%) of the variance in the fluency measure is between candidates, and half is variance within candidates across test occasions. The intercept of 290.53 in Model 1 is simply the average number of words per script across all candidates and occasions (see Table 8). The intercept variance (2235.96) was significant (X^2 = 306.51, *df*.= 77, *p*<.01), indicating, not surprisingly, that the average number of words per script varied significantly across candidates.

Model 2 added occasion as a linear predictor at level 1; the relation between occasion and fluency (i.e., rate of change in fluency) was allowed to vary across candidates. The model predicts a value of 269.89 words at test occasion 1 (i.e., average number of words per script across all candidates at test occasion 1; see Table 8), which increases by 21.64 words on average on each succeeding test occasion. This increase is statistically significant. Additionally, the occasion slope variance (450.97) was significant (X^2 = 128.61, *df*.= 77, *p*<.01), indicating that the rate of change in fluency across test occasion varied significantly across candidates. Fit statistics indicated that Model 2 fits the data significantly better than Model 1 (X^2 = 44.18, *df*.= 2, *p*<.01). Model 3 added the time-invariant predictor, candidate group (i.e., time 1 writing score), at level 2.

As Table 9 shows, the relationship between candidate group and fluency was significant. For each one-point increase in writing score at time 1, there is a significant increase of 29.95 words, on average, in script length. To explain the variance in the rate of change in fluency over test occasions across candidates, Model 4 included cross-level interactions with occasion. Specifically, the time-invariant predictor candidate group was added in order to estimate the relationship between candidate initial writing score and the rate of change in fluency across test occasions. The last column of Table 9 shows that candidate group had a negative effect on the rate of change in fluency across occasion (-8.87), but this effect was not significant.

Based on the results above, the final model for fluency is Model 3 in Table 9. This model includes one predictor at level 1, test occasion, and one predictor at level 2, candidate group (i.e., time 1 writing score). It specifies changes in fluency as a function of test occasion and differences in fluency at time 1 as a function of differences in candidate L2 writing abilities at time 1 (i.e., time 1 writing score).

Fit statistics indicated that Model 3 fits the data significantly better than Model 1 (X^2 = 68.33, *df*.= 2, *p*<.01). According to Model 3, the average number of words per script for candidates scoring 4 on Writing Task 2 at time 1 was 230.67 words. For each one-band increase in writing scores at time 1, there is a significant increase in script length by 38.23 words, on average. Additionally, there was a significant increase of script length by 30.42 words, on average, on each succeeding test occasion. The final model accounted for ([2250.47-1345.49]/2250.47 =) 40% of the variance in fluency across test occasions and ([3214.04-2305.16]/3214.04=) 28% of the variance between candidates.

As the high and significant variance coefficients for the intercept (2214.04) and the occasion slope (451.12) indicate, much of the variance in fluency between (72%) and within (60%) candidates is not explained by the final model. Other candidate factors and covariates may explain the remaining variance in fluency.

3.2.2 Linguistic accuracy

Table 10 displays descriptive statistics concerning the ratio of each error type across test occasions and candidate groups. It shows that the candidates made, on average, fewer errors at test occasion 3 (*M*= 13 errors per 100 words) than they did at test occasion 2 (*M*= 15) and test occasion 1 (*M*= 16). This pattern of fewer errors on subsequent occasions compared to preceding ones seems to apply to some of the error types as well (e.g., style). Furthermore, candidates who scored 4 at test occasion 1 made more errors (*M*= 20 errors per 100 words) than did those scoring 5 (*M*= 14), who in turn made more errors than did those who scored 6 (*M*= 10) at test occasion 1.

Additionally, for all error types, candidates scoring 4 at test occasion 1 made more errors per 100 words than did those scoring 5 and 6.

The autocorrelations (Pearson *r*) of the accuracy measures over time (Table 11) were positive indicating that, generally, candidates who made more errors at each test occasion made more errors at the following test occasion and vice versa. This was particularly true for mechanics errors (see Table 11).

These findings indicate that the order of the candidates relative to each other in terms of accuracy was somewhat stable across test occasions.

Candidate group	4		5		6		Total	
Test occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1								
All errors	21	8	14	6	11	6	16	8
Grammar	2	1	1	1	1	1	1	1
Usage	3	2	2	2	2	1	2	1
Mechanics	6	4	3	2	3	2	4	3
Style	10	7	7	5	5	5	8	6
Occasion 2								
All errors	20	7	16	6	10	4	15	7
Grammar	2	1	1	1	1	1	1	1
Usage	3	1	3	2	2	1	2	1
Mechanics	6	4	3	2	2	1	4	3
Style	10	6	9	6	5	4	8	6
Occasion 3								
All errors	19	7	13	5	8	4	13	7
Grammar	2	1	1	1	1	0	1	1
Usage	3	1	2	1	2	1	2	1
Mechanics	6	3	3	2	3	1	4	3
Style	9	6	7	4	4	4	7	5
Total								
All errors	20	7	14	6	10	5	15	8
Grammar	2	1	1	1	1	1	1	1
Usage	3	1	2	1	2	1	2	1
Mechanics	6	4	3	2	3	1	4	3
Style	10	6	8	5	5	4	7	6

Table 10: Descriptive statistics for linguistic accuracy by candidate group and test occasion

	Occasions 1 and 2	Occasions 2 and 3
All errors	$.55^{**}$	$.57^{**}$
Grammar	$.24^{*}$	$.29^{*}$
Usage	$.42^{**}$	$.42^{**}$
Mechanics	$.82^{**}$	$.76^{**}$
Style	$.30^{**}$	$.36^{**}$

* *p*<.05; ** *p*<.01; N=78

Table 11: Autocorrelations for accuracy measures

Table 12 displays the MLM results for accuracy; only the ratio of the total number of errors was examined. The results for Model 1 indicate that intra-individual variability (26.60) and inter-individual variability (29.77) in accuracy were almost equal. ICC is .53, indicating that slightly more than half (53%) of the variance in accuracy is between candidates and 47% is variance within candidates across test occasions. The intercept of 14.74 in Model 1 indicates that the average number of errors across all candidates and occasions is about 15 errors per 100 words (see Table 10). The intercept variance (335.51) was significant indicating that the average number of errors varied significantly across candidates. The results for Model 2 show that the average number of errors across all candidates at test occasion 1 was 15.78, which decreases by 1.03 errors per 100 words, on average, on each succeeding test occasion. This decrease is statistically significant. However, the rate of change in accuracy did not vary significantly across candidates. Model 3 indicated that there were significant differences between candidate groups in terms of accuracy at time 1, while Model 4 indicated that candidate group did not have a significant effect on the rate of change in accuracy across test occasions.

	Model 1	Model 2	Model 3
Fixed effects (SE)			
Intercept Candidate Group Occasion	14.74** (.70)	15.78** (.84) -1.03* (.42)	21.14** (1.02) -5.36** (.63) -1.03* (.42)
Random effects			
Between-candidate X^2 (df)	29.77 335.51** (77)	37.10 223.55** (77)	19.63 153.59** (76)
Occasion slope X^2 (df)		2.33 92.32 (77)	2.34 92.37 (77)
Within-candidate	26.60	23.39	23.38
Model fit			
Deviance (#parameters)	1542.62 (2)	1536.88 (4)	1480.03 (4)
Model comparison: X^2 (df.)		5.74* (2)	62.59** (2)

* *p*<.05; ** *p*<.01; N=78

Table 12: MLM results for linguistic accuracy

Based on the results from Models 1-4, the final model for accuracy is Model 3 which includes test occasion and one time-invariant predictor at level 2, candidate group (i.e., time 1 writing score). This model specifies changes in accuracy as a function of test occasion; differences in accuracy at test occasion 1 are specified as a function of differences in candidate initial L2 writing ability (i.e., time 1 writing score). As Table 12 shows, fit statistics indicated that Model 3 fits the data significantly better than Model 1 ($X^2 = 62.59$, *df.*= 2, *p*<.01). According to Model 3, the average number of errors per 100 words for candidates with writing score 4 at test occasion 1 was 21.14 errors. For each one-band increase in writing scores at test occasion 1, there was a significant decrease in the number of errors by 5.36 errors (per 100 words), on average. Additionally, there was a significant decrease in the number of errors by 1.03 errors per 100 words, on average, on each succeeding test occasion. The rate of change in accuracy did not vary significantly across candidates. Model 3 explained 47% of the between-person variance and 12% of the within-person variance. The variance for the intercept (19.63) was significant indicating that much of the variance in accuracy between candidates is not explained by the final model. Other candidate factors may explain the remaining variance.

3.2.3 Syntactic complexity

Table 13 displays descriptive statistics for each of the three measures of syntactic complexity across test occasions and candidate groups. There does not seem to be much difference in any of the measures across candidate groups or test occasions, with the exception, perhaps, of NP density which seems to vary across candidate groups. For example, candidates scoring 6 at test occasion 1 seem to have a higher NP density index (*M*= .84) than do those scoring 4 (*M*= .73) and 5 (*M*= 76).

The autocorrelations (Pearson *r*) of the complexity measures across test occasions (Table 14) tended to be positive indicating that, generally, candidates who had higher levels of syntactic complexity at each test occasion had higher levels of syntactic complexity at the following test occasion and vice versa. This was particularly true for mean sentence syntactic similarity, but less so for left embeddedness for test occasions 1 and 2 and NP density for test occasions 2 and 3 (see Table 14).

Candidate group	**4**		**5**		**6**		**Total**	
Test occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1								
Left embeddedness	4.74	2.62	4.19	1.14	4.83	1.57	4.59	1.88
Syntactic similarity	0.10	0.03	0.11	0.03	0.10	0.02	0.10	0.03
NP density	0.72	0.12	0.77	0.16	0.83	0.17	0.78	0.15
Occasion 2								
Left embeddedness	4.34	1.74	4.16	1.22	5.33	1.62	4.61	1.61
Syntactic similarity	0.11	0.03	0.10	0.03	0.10	0.02	0.10	0.03
NP density	0.74	0.14	0.76	0.10	0.84	0.16	0.78	0.14
Occasion 3								
Left embeddedness	4.61	1.45	4.69	1.17	4.6	1.21	4.63	1.27
Syntactic similarity	0.11	0.03	0.10	0.03	0.10	0.02	0.10	0.03
NP density	0.71	0.13	0.75	0.17	0.87	0.13	0.78	0.16
Total								
Left embeddedness	4.56	1.98	4.35	1.19	4.92	1.49	4.61	1.60
Syntactic similarity	0.11	0.03	0.10	0.03	0.10	0.02	0.10	0.03
NP density	0.73	0.13	0.76	0.14	0.84	0.15	0.78	0.15

Table 13: Descriptive statistics for syntactic complexity by candidate group and test occasion

	Occasions 1 and 2	**Occasions 2 and 3**
Left embeddedness	.06	$.38^{**}$
NP density	$.29^{**}$	.19
Syntax similarity	$.40^{**}$	$.48^{**}$

* *p*<.05; ** *p*<.01; N=78

Table 14: Autocorrelations for syntactic complexity measures

	Left embeddedness		**Syntax similarity**	
	Model 1	**Model 2**	**Model 1**	**Model 2**
Fixed effects (SE)				
Intercept Occasion	4.61** (.11)	4.59** (.19) .02 (.14)	.10** (.002)	.10** (.003) -.0003 (.001)
Random effects				
Between-candidate X^2 (df)	.18 94.14 (77)	1.43 151.35** (77)	.0003 253.59** (77)	.0004 178.43** (77)
Time slope X^2 (df)		.61 130.06** (77)		.00004 91.81 (77)
Within-candidate	2.38	1.78	.0004	.0004
Model fit				
Deviance (#parameters)	882.00 (3)	878.22 (4)	-1060.87 (3)	-1049.20 (4)
Model comparison: X^2 (df.)		3.78 (2)		11.68** (2)

* *p*<.05; ** *p*<.01; N=78

Table 15: MLM results for left embeddedness and syntax similarity

MLM analyses were conducted for each of the three complexity measures separately. Table 15 reports the MLM results for left embeddedness. Model 1 results indicated that most of the variance (2.38) was within candidate. Between-person variance (.16) was not significant and accounted for only 6% of the total variance in left embeddedness. This means that the differences between the candidates in this study in terms of left embeddedness indices were not statistically significant. Model 2, which included test occasion, indicated that left embeddedness increased by .02 on average on each succeeding occasion, but this increase was not statistically significant. However, the rate of change in left embeddedness over time varied significantly across candidates (X^2 = 130.06, *df*.= 77, *p*<.01). Models 3 and 4 indicated that (a) there were no significant differences between candidate groups in terms of left embeddedness at test occasion 1 and (b) candidate group did not have a significant effect on the rate of change in left embeddedness across test occasions.

Consequently, the final model for left embeddedness is Model 2, which includes only test occasion, but allows the rate of change in left embeddedness over time to vary across candidates. Model 2 shows that the average left embeddedness index for all candidates at time 1 was 4.59 and that there was a non-significant increase in left embeddedness of .02, on average, on each succeeding test occasion.

However, the rate of change in left embeddedness across test occasions varied significantly across candidates. The inclusion of test occasion explained only 25% of within-individual variance. However, none of the between-person variance in change rate across test occasions was explained by the model.

Table 15 reports the MLM results for syntax similarity as well. Model 1 results indicated that 43% of the variance in syntax similarity was between candidates (.0003). Between-individual variance was significant indicating that the difference between candidates in terms of syntax similarity was statistically significant. Model 2 indicated that syntax similarity decreased by .0003 on average on each succeeding test occasion, but this decrease was not statistically significant. Nor did the rate of change in syntax similarity vary significantly across candidates. Model 3 indicated that there were no significant differences between candidate groups in terms of syntax similarity at time 1. Consequently, the final model for syntax similarity is Model 1 in Table 15.

Finally, Table 16 reports the MLM results for NP density. The results for Model 1 indicated that 23% of the variance in NP density indices (.006) was between candidates; the remaining variance was within candidates. The variance between candidates was significant indicating that NP density varied significantly across candidates.

Model 2 indicated that NP density increased by .002 on average on each succeeding test occasion, but this increase was not statistically significant. Nor did the rate of change in NP density vary significantly across candidates (X^2 = 73.82, *df*.= 77, *p*>.05). Model 3, however, indicated that there were significant differences between candidate groups in terms of NP density at test occasion 1. As a result, the final model for NP density is Model 3 which included test occasion at level 1 and candidate group at level 2.

As the last column of Table 16 shows, the average NP density for candidates with writing score of 4 at test occasion 1 was .71. There was a non-significant increase in NP density by .002, on average, on each succeeding test occasion. Candidate group was significantly associated with NP density. Specifically, for each increase of one band in writing scores at test occasion 1, NP density increased by .06.

While the rate of change in NP density did not vary significantly across candidates, the between-person variance (.006) was statistically significant, indicating that candidate group did not explain all the variance between candidates in terms of NP density. The final model explained only 14% of the between-person variance. None of the within-person variance in NP density was explained by Model 3.

	Model 1	Model 2	Model 3
Fixed effects (SE)			
Intercept	.78** (.01)	.78** (.02)	.71** (.02)
Candidate group			.06** (.01)
Occasion		.002 (.01)	.002 (.01)
Random effects			
Between-candidate	.006	.007	.006
X^2 (df)	170.09** (77)	114.84** (77)	108.51** (76)
Time slope		.00004	.0002
X^2 (df)		73.82 (77)	74.42 (77)
Within-candidate	.02	.02	.02
Model fit			
Deviance (#parameters)	-238.52 (2)	-229.41 (4)	-242.35 (4)
Model comparison: X^2 (df.)		9.11* (2)	3.83 (2)

* *p*<.05; ** *p*<.01; N=78

Table 16: MLM results for NP density

3.2.4 Lexical features

Table 17 displays descriptive statistics for the four lexical measures across test occasions and candidate groups. It shows that, while the three candidate groups seem to differ in terms of the four indices, none of the four indices seems to vary across test occasions. For example, candidates scoring 6 at test occasion 1 had higher MTLD indices (*M*= 87.19), than did those scoring 4 (*M*= 65.40) and 5 (*M*= 72.16). Furthermore, candidates scoring 6 at test occasion 1, generally, used longer words (*M*= 4.98 letters per word) and more low-frequency words (*M*= 2.33) than did candidates scoring 4 and 5 as shown in Table 17.

The patterns in Table 17 suggest that candidates scoring 6 at test occasion 1 tended to use more content words (i.e., higher lexical density), more diverse vocabulary, longer words, and more low-frequency vocabulary than did candidates with lower initial writing scores (4 and 5).

The autocorrelations (Pearson *r*) of the lexical measures across test occasions are reported in Table 18. Table 18 shows that the correlations are positive and significant indicating that, generally, candidates with higher indices on each of the four lexical measures at each test occasion had higher indices on that measure at the following test occasion and vice versa. This is particularly the case for AWL and word frequency.

Table 19 displays the MLM results for lexical density. The results for Model 1 indicated that there was the same intra-individual variability as inter-individual variability (.001) in lexical density. That is, half of the variance in lexical density is between candidates. The intercept variance (.001) was significant (X^2 = 203.93, df.= 77, p<.01), indicating that lexical density varied significantly across candidates. The results for Model 2 indicated that lexical density increased by .001 on average on each succeeding test occasion, but this increase was not statistically significant. However, the rate of change in lexical density over time varied significantly across candidates (X^2 = 113.20, *df*.= 77, *p*<.01). Furthermore, Models 3 and 4 indicated that (a) there was a significant effect of candidate group on lexical density at test occasion 1, but (b) candidate group did not have a significant effect on the rate of change in lexical density across test occasions.

Consequently, the final model for lexical density is Model 3. As the last column of Table 19 shows, Model 3 predicts that the average lexical density for all candidates scoring 4 at test occasion 1 was .51. For each one-band increase in initial writing scores, there was a significant increase in lexical density by .01, on average. The change in lexical density over time (.001) was not significant, but the rate of change in lexical density over time varied significantly across candidates (X^2 = 113.22, *df*.= 77, *p*<.01). Although it fit the data significantly better than Model 1 (X^2 = 8.79, *df*.= 2, *p*<.01), Model 3 explained none of the within-person or between-person variance. Nor did it explain any of the variance in the rate of change in lexical density across test occasions.

Candidate group	**4**		**5**		**6**		**Total**	
Test occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1								
Lexical density	.52	.05	.51	.04	.54	.03	.52	.04
MTLD,	64.26	16.21	69.98	19.06	86.68	20.57	73.64	20.79
AWL	4.49	0.28	4.63	0.31	5.06	0.24	4.73	0.37
Word frequency	2.53	0.13	2.49	0.13	2.33	0.11	2.45	0.15
Occasion 2								
Lexical density	.52	.05	.53	.03	.54	.05	.53	.04
MTLD,	65.75	17.13	70.18	15.66	84.95	18.92	73.63	18.96
AWL	4.59	0.36	4.61	0.25	4.9	0.26	4.7	0.32
Word frequency	2.49	0.12	2.5	0.10	2.33	0.14	2.44	0.14
Occasion 3								
Lexical density	.52	.05	.52	.04	.54	.04	.53	.04
MTLD,	66.19	17.83	76.31	16.55	89.94	19.21	77.48	20.2
AWL	4.55	0.33	4.64	0.32	4.99	0.23	4.73	0.35
Word frequency	2.55	0.14	2.47	0.12	2.32	0.13	2.45	0.16
Total								
Lexical density	.52	.05	.52	.04	.54	.04	.53	.04
MTLD,	65.4	16.87	72.16	17.18	87.19	19.43	74.92	19.99
AWL	4.55	0.32	4.63	0.29	4.98	0.25	4.72	0.34
Word frequency	2.52	0.13	2.49	0.12	2.33	0.13	2.45	0.15

Table 17: Descriptive statistics for lexical measures by candidate group and test occasion

	Occasions 1 and 2	**Occasions 2 and 3**
Lexical density	.39**	.45**
MTLD	.53**	.41**
AWL	.68**	.66**
Frequency	.60**	.57**

* *p*<.05; ** *p*<.01; N=78

Table 18: Autocorrelations for lexical measures

	Model 1	**Model 2**	**Model 3**
Fixed effects (SE)			
Intercept	.53** (.004)	.53** (.005)	.51** (.007)
Candidate group			.01* (.004)
Occasion		.001 (.003)	.001 (.003)
Random effects			
Between-candidate	.001	.001	.001
X^2 (df)	203.93** (77)	158.06** (77)	149.89** (76)
Occasion slope		.0002	.0002
X^2 (df)		113.20** (77)	113.22** (77)
Within-candidate	.001	.001	.001
Model fit			
Deviance (#parameters)	-822.95 (3)	-814.47 (4)	-814.16 (4)
Model comparison: X^2 (df.)		8.48** (2)	8.79** (2)

* *p*<.05; ** *p*<.01; N=78

Table 19: MLM results for lexical density

Table 20 displays the MLM results for lexical variation (MTLD). The results for Model 1 indicated that there was almost the same intra-individual variability (205.27) as inter-individual variability (192.71) in lexical variation. Specifically, 48% of the variance in lexical variation is between candidates. The between-person variance was significant (X^2 = 297.68, df.= 77, p<.01), indicating that lexical variation varied significantly across candidates. The results for Model 2 indicated that lexical variation increased by 1.92 on average on each succeeding test occasion, but this increase was not statistically significant. Nor did the rate of change in MTLD vary significantly across candidates (X^2 = 74.81, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on MTLD at test occasion 1.

Consequently, the final model for MTLD is Model 3. As the last column of Table 20 shows, according to Model 3, the average MTLD for all candidates scoring 4 at test occasion 1 was 62.05. For each one-band increase in initial writing scores, there was a significant increase in MTLD by 10.94, on average. The change in MTLD across test occasions (1.92) was not significant; nor did the rate of change in MTLD density across test occasions vary significantly across candidates (X^2 = 75.17, *df*.= 77, *p*>.05). Model 3 fits the data significantly better than Model 1 (X^2 = 36.72, *df*.= 2, *p*<.01). Additionally, it explains 32% of the between-person variance and 2% of the within-person variance in lexical variation.

Table 21 displays the MLM results for average word length (AWL). The results for Model 1 indicated that most of the variance (.08 or 67%) was between candidates. Intra-individual variability (.04) accounted for only 33% of the variance in AWL. The intercept variance (.08) was significant (X^2 = 526.05, df.= 77, p<.01), indicating that AWL varied significantly across candidates. The results for Model 2 indicated that AWL increased by .0005 on average on each succeeding test occasion, but this increase was not statistically significant. Nor did the rate of change in AWL vary significantly across candidates (X^2 = 96.23, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on AWL at test occasion 1.

Consequently, the final model for AWL is Model 3. As the last column of Table 21 shows, fit statistics indicate that Model 3 fits the data significantly better than Model 1 (X^2 = 22.99, *df*.= 2, *p*<.01). The results for Model 3 show that the average AWL for candidates with a writing score of 4 at test occasion 1 was 4.50 letters per word and that there was a non-significant increase in AWL by .0005 letters, on average, on each succeeding test occasion. Furthermore, candidate group was significantly associated with AWL. Specifically, for each increase of one band in initial writing scores, AWL increased by .22 letters. The rate of change in AWL over time did not vary significantly across candidates (X^2 = 96.24, *df*.= 77, *p*>.05). The final model explained 44% of the between-person variance, but no within-person variance in AWL.

	Model 1	**Model 2**	**Model 3**
Fixed effects (SE)			
Intercept	74.92** (1.83)	72.99** (2.22)	62.05** (2.73)
Candidate group			10.94** (1.85)
Occasion		1.92 (1.11)	1.92 (1.11)
Random effects			
Between-candidate	196.10	224.50	153.44
X^2 (df)	297.68** (77)	178.15** (77)	144.25** (76)
Occasion slope		1.91	2.82
X^2 (df)		74.81 (77)	75.17 (77)
Within-candidate	205.27	201.08	200.14
Model fit			
Deviance (#parameters)	2009.54 (2)	2006.03 (4)	1972.82 (4)
Model comparison: X^2 (df.)		3.51 (2)	36.72** (2)

* *p*<.05; ** *p*<.01; N=78

Table 20: MLM results for lexical variation

	Model 1	Model 2	Model 3
Fixed effects (SE)			
Intercept	4.72** (.03)	4.72** (.04)	4.50** (.05)
Candidate group			.22** (.03)
Occasion		.0005 (.02)	.0005 (.02)
Random effects			
Between-candidate	.08	.09	.05
X^2 (df)	526.05** (77)	306.73** (77)	200.50** (76)
Occasion slope		.005	.005
X^2 (df)		96.23 (77)	96.24 (77)
Within-candidate	.04	.04	.04
Model Fit			
Deviance (#parameters)	66.62 (2)	73.63 (4)	43.63 (4)
Model comparison: X^2 (df.)		7.01* (2)	22.99** (2)

* *p*<.05; ** *p*<.01; N=78

Table 21: MLM results for AWL

Table 22 displays the MLM results for word frequency. The results for Model 1 indicated that about two-thirds of the variance (.014 or 61%) was between-candidates. Intra-individual variability (.009) accounted for 39% of the variance in word frequency. The intercept variance (.014) was significant (X^2 = 460.19, df.= 77, p<.01), indicating that word frequency varied significantly across candidates. The results for Model 2 indicated that word frequency decreased by .002 on average on each succeeding test occasion, but this decrease was not statistically significant. Nor did the rate of change in word frequency vary significantly across candidates (X^2 = 66.85, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on word frequency at time 1. Consequently, the final model for word frequency is Model 3. As the last column of Table 22 shows, fit statistics indicate that Model 3 fits the data significantly better than Model 1 (X^2 = 22.58, *df*.= 2, *p*<.01). According to Model 3, the average word frequency for candidates with a writing score of 4 at test occasion 1 was 2.55. The decrease in word frequency (-.002) across test occasions was not significant; nor did the rate of change in word frequency across test occasions vary significantly across candidates (X^2 = 66.77, *df*.= 77, *p*>.05). However, candidate group was significantly associated with word frequency. Specifically, for each increase of 1 band in initial writing scores, word frequency decreased by .10. Model 3 explained 20% of the between-person variance, but none of the within-person variance in word frequency.

	Model 1	Model 2	Final Model
Fixed effects (SE)			
Intercept	2.45** (.01)	2.45** (.02)	2.55** (.02)
Candidate group			-.10** (.01)
Occasion		-.002 (.007)	-.002 (.007)
Random effects			
Between-candidate	.014	.01	.008
X^2 (df)	460.19** (77)	214.86** (77)	155.44** (76)
Occasion slope		.00002	.00001
X^2 (df)		66.85 (77)	66.77 (77)
Within-candidate	.009	.009	.009
Model fit			
Deviance (#parameters)	-305.57 (2)	-296.00 (4)	-328.15 (4)
Model comparison: X^2 (df.)		9.57** (2)	22.58** (2)

* *p*<.05; ** *p*<.01; N=78

Table 22: MLM results for word frequency

3.2.5 Coherence and cohesion

Table 23 displays descriptive statistics for each of the four measures of cohesion and coherence across test occasions and candidate groups. The results in Table 23 do not reveal any large differences across candidate groups or test occasions. The autocorrelations (Pearson *r*) of the coherence measures over time (Table 24) were positive and significant indicating that, generally, candidates who had higher indices on each of the four coherence and cohesion measures at each test occasion had higher indices at the following test occasion and vice versa.

Candidate group	**4**		**5**		**6**		**Total**	
Test occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1								
All connectives	102.81	22.1	105.39	20.70	103.49	17.68	103.9	20.01
Argument overlap	0.55	0.23	0.55	0.17	0.55	0.16	0.55	0.19
LSA overlap, sentences	0.20	0.09	0.18	0.07	0.19	0.05	0.19	0.07
LSA overlap, paragraphs	0.32	0.16	0.40	0.16	0.41	0.09	0.38	0.14
Occasion 2								
All connectives	107.16	23.25	103	18.87	101.06	18.60	103.74	20.25
Argument overlap	0.60	0.21	0.55	0.20	0.54	0.12	0.56	0.18
LSA overlap, sentences	0.19	0.08	0.20	0.10	0.20	0.05	0.20	0.08
LSA overlap, paragraphs	0.38	0.17	0.43	0.18	0.44	0.13	0.42	0.16
Occasion 3								
All connectives	102.93	16.61	108.94	22.24	100.98	16.94	104.28	18.84
Argument overlap	0.64	0.21	0.64	0.19	0.51	0.15	0.60	0.19
LSA overlap, sentences	0.20	0.08	0.20	0.07	0.19	0.06	0.20	0.07
LSA overlap, paragraphs	0.43	0.13	0.42	0.13	0.40	0.12	0.41	0.13
Total								
All connectives	104.3	20.68	105.77	20.53	101.84	17.56	103.97	19.62
Argument overlap	0.60	0.22	0.58	0.19	0.53	0.14	0.57	0.19
LSA overlap, sentences	0.20	0.08	0.20	0.08	0.19	0.05	0.20	0.07
LSA overlap, paragraphs	0.38	0.16	0.42	0.16	0.42	0.12	0.40	0.15

Table 23: Descriptive statistics for cohesion and coherence measures by candidate group and test occasion

	Occasions 1 and 2	**Occasions 2 and 3**
All connectives	.38**	.31**
Argument overlap	.40**	.27*
LSA overlap, paragraphs	.25*	.45**
LSA overlap, sentences	.31**	.44**

* *p*<.05; ** *p*<.01; N=78

Table 24: Autocorrelations for coherence and cohesion measures

MLM analyses indicated that the results for connectives density and argument overlap for adjacent sentences were similar. Specifically, for each index: (a) there was no significant change across test occasions; (b) there was no significant effect of candidate group on the index at test occasion 1; (c) the rate of change in the index across test occasions did not vary significantly across candidates; and (d) candidate group did not have a significant effect on the rate of change in the index across test occasions.

Consequently, the final model for each of the three indices is Model 1 in Table 25. For example, the results for Model 1 indicated that most of the variance in connectives density (235.10 or 61%) was within candidates. Inter-individual variability (148.39) accounted for 39% of the variance in connectives incidence. The intercept variance was significant (X^2 = 225.69, df.= 77, p<.01), indicating that connectives density varied significantly across candidates. The results for Model 2 indicated that connectives density increased by .19 (or almost 2 connectives per 100 words) on average on each succeeding test occasion, but this increase was not statistically significant. Nor did the rate of change in connectives density vary significantly across candidates (X^2 = 64.46, *df*.= 77, *p*>.05). For argument overlap for adjacent sentences, most of the variance (.02 or 67%) was within candidates. The intercept variance, though small (.01), was significant (X^2 = 182.30, df.= 77, p<.01), indicating that argument overlap varied significantly across candidates. The results for Model 2 indicated that argument overlap increased by .02 on average on each succeeding test occasion, but this increase was not statistically significant. The rate of change in argument overlap did not vary significantly across candidates (X^2 = 90.41, *df*.= 77, *p*>.05).

As for mean LSA overlap for adjacent sentences, Table 25 shows that most of the variance (.004 or 77%) was within candidates. The intercept variance, though small (.002), was significant (X^2 = 180.49, df.= 77, p<.01), indicating that mean LSA overlap for adjacent sentences varied significantly across candidates. The results for Model 2 indicated that mean LSA overlap for adjacent sentences increased by .004 on average on each succeeding test occasion, but this increase was not statistically significant. However, the rate of change in LSA overlap for adjacent sentences varied significantly across candidates (X^2 = 101.98, *df*.= 77, *p*<.05).
The inclusion of test occasion in Model 2 explained 25% of the within-person variance in mean LSA overlap for adjacent sentences. Fit statistics indicate that Model 2 fits the data significantly better than Model 1 (X^2 = 8.20, *df*.= 2, *p*<.05).

Finally, for mean LSA overlap for adjacent paragraphs, Table 26 shows that most of the variance (.02 or 77%) was within candidates. The intercept variance, though small (.006), was significant (X^2 = 161.06, df.= 77, p<.01), indicating that mean LSA overlap for adjacent paragraphs varied significantly across candidates. The results for Model 2 indicated that mean LSA overlap for adjacent paragraphs increased by .02 on average on each succeeding test occasion, but this increase was not statistically significant. Nor did the rate of change in mean SLA overlap for adjacent paragraphs over time vary significantly across candidates (X^2 = 85.23, *df*.= 77, *p*>.05). However, Models 3 and 4 indicated that (a) there was a significant effect of candidate group on mean LSA overlap for adjacent paragraphs at test occasion 1 and (b) candidate group had a significant effect on the rate of change in mean LSA overlap for adjacent paragraphs over time.

Consequently, the final model for lexical density is Model 4. As the last column of Table 26 shows, according to Model 4, the average mean LSA overlap for adjacent paragraphs for candidates with writing score 4 at test occasion 1 was .34. For each one-band increase in writing scores at time 1, there was a significant increase in mean LSA overlap for adjacent paragraphs by .05, on average. Additionally, there was a significant increase in mean LSA overlap for adjacent paragraphs by .05, on average, on each succeeding test occasion. The rate of change in mean LSA overlap for adjacent paragraphs, however, was moderated by a significant effect for candidate group.

Overall, the rate of change in mean LSA overlap for adjacent paragraphs was weaker by .03, on average, for each one-band increase in initial writing scores. This means that candidates with higher initial writing scores exhibited a lower rate of change in mean LSA overlap for adjacent paragraphs compared to candidates with lower initial writing scores. Fit statistics indicate that Model 4 fits the data significantly better than Model 1 (X^2 = 11.94, *df*.= 2, *p*<.05). Model 4 explained 17% of the between-person variance and 50% of the within-person variance in mean LSA overlap for adjacent paragraphs.

	Connectives density		Argument overlap		LSA sentence	
	Model 1	**Model 2**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
Fixed effects (SE)						
Intercept	103.97** (1.71)	103.78** (2.12)	.57** (.02)	.55** (.02)	.20** (.006)	.19** (.008)
Occasion		.19 (1.12)		.02 (.01)		.004 (.005)
Candidate group						
Random effects						
Between-candidate	151.34	177.43	.01	.01	.002	.002
X^2 (df= 77)	225.69**	139.48**	182.30**	137.68**	180.49**	136.34**
Occasion slope		1.36		.002		.0005
X^2 (df)		64.46 (77)		90.41 (77)		101.98* (77)
Within-candidate	235.10	235.26	.02	.02	.004	.003
Model fit						
Deviance (#parameters)	2019.83 (2)	2018.96 (4)	-131.52 (2)	-126.67 (4)	-575.95 (2)	-567.75 (4)
Model comparison: X^2 (df.)		.87 (2)		4.85 (2)		8.20* (2)

* *p*<.05; ** *p*<.01; N=78

Table 25: MLM results for connectives density, argument overlap, and mean LSA for adjacent sentences

	Model 1	**Model 2**	**Model 4**
Fixed effects (SE)			
Intercept	.40** (.01)	.39** (.02)	.34** (.03)
Candidate group			.05** (.02)
Occasion		.02 (.01)	.05** (.02)
Candidate group			-.03** (.01)
Random effects			
Between-candidate	.006	.006	.005
X^2 (df= 77)	161.06**	119.81**	109.72** (76)
Occasion slope		.0008	.0003
X^2 (df)		85.23 (77)	97.15 (76)
Within-candidate	.02	.01	.01
Model fit			
Deviance (#parameters)	-245.46 (2)	-239.48 (4)	-233.51 (4)
Model Comparison: X^2 (df.)		5.99* (2)	11.94** (2)

* *p*<.05; ** *p*<.01; N=78

Table 26: MLM results for mean LSA for adjacent paragraphs

3.2.6 Discourse structure

Criterion was used to examine script organisation, i.e., whether each script included each of five discourse elements (introductory material, thesis statement, main idea, supporting ideas, and conclusion), and development (i.e., the percentage of the script assigned to each discourse element included in the script). Tables 27 and 28 display descriptive statistics for script organisation and development, respectively, across test occasions and candidate groups. All candidates included supporting ideas in their scripts at all test occasions. The percentage of candidates who included other discourse elements varied across candidate groups and test occasions.

To examine whether there are significant associations between test occasion and candidate group, on the one hand, and the presence or absence of each discourse element, on the other, Chi-square (X^2) tests were conducted for each discourse element separately. The results indicated that there was a significant association of candidate group with introduction (X^2= 6.04, *df*.= 2 *p*<.05) and conclusion (X^2= 7.32, *df*.= 2 *p*<.05), but not for the other discourse elements. As Table 27 shows, a significantly higher proportion of candidates scoring 6 at test occasion 1 included a conclusion (92%) and an introduction (68%) than did candidates scoring 5 (86% and 60%, respectively) and 4 (77% and 49%, respectively). There was no significant association between test occasion and the presence of any of the discourse elements.

As for development (i.e., length of each discourse element), Table 28 does not show any large differences across test-occasions or candidate groups. It seems that the relative length of each of the five discourse elements was stable across candidate groups and test occasions. MLM results (not reported here) indicated that for all five discourse elements: (a) the intercept, that is the relative length, of each discourse element at test occasion 1 varied significantly across candidates; (b) there was no significant change in length across test occasions; (c) the rate of change in length across test occasions did not vary significantly across candidates; (d) candidate group did not have a significant effect on the rate of change in length across test occasions; and (e) there was no significant effect of candidate group on length at test occasion 1, except for conclusion. MLM results indicated that the average length of the conclusion for candidates with writing score 4 at test occasion 1 was 9.88%. For each one-band increase in initial writing scores, there was a significant increase in the length of the conclusion by 1.40%, on average.

Candidate group	4		5		6		Total	
	f	%	f	%	f	%	f	%
Occasion 1								
Introduction	10	38	15	58	19	73	44	56
Thesis	21	81	23	88	18	69	62	79
Main idea	25	96	25	96	26	100	76	97
Support	26	100	26	100	26	100	78	100
Conclusion	18	69	22	85	24	92	64	82
Occasion 2								
Introduction	14	54	16	62	16	62	46	59
Thesis	22	85	20	77	20	77	62	79
Main idea	25	96	25	96	26	100	76	97
Support	26	100	26	100	26	100	78	100
Conclusion	18	69	22	85	24	92	64	82
Occasion 3								
Introduction	14	54	16	62	18	69	48	62
Thesis	21	81	21	81	21	81	63	81
Main idea	26	100	25	96	26	100	77	99
Support	26	100	26	100	26	100	78	100
Conclusion	24	92	23	88	24	92	71	91
Total								
Introduction	38	49	47	60	53	68	138	59
Thesis	64	82	64	82	59	76	187	80
Main idea	76	97	75	96	78	100	229	98
Support	78	100	78	100	78	100	234	100
Conclusion	60	77	67	86	72	92	199	85

Table 27: Descriptive statistics for organisation by candidate group and test occasion

Candidate group	4		5		6		Total	
	M	SD	M	SD	M	SD	M	SD
Occasion 1								
Introduction	3.57	5.26	7.65	7.99	8.00	7.36	6.40	7.17
Thesis	11.16	8.82	11.57	7.46	8.27	7.64	10.33	8.03
Main idea	20.33	14.41	17.18	12.59	18.33	9.09	18.61	12.15
Support	51.18	19.29	48.21	14.96	50.56	12.30	49.98	15.63
Conclusion	9.07	7.55	13.57	10.60	12.73	6.01	11.79	8.40
Occasion 2								
Introduction	5.95	6.96	10.11	11.21	6.32	7.41	7.46	8.83
Thesis	9.17	6.60	8.14	7.72	9.08	7.47	8.80	7.20
Main idea	15.35	8.90	15.38	9.26	19.56	11.88	16.76	10.17
Support	56.31	15.50	51.02	13.27	51.82	9.74	53.05	13.10
Conclusion	10.60	9.01	12.26	7.48	12.07	5.30	11.64	7.36
Occasion 3								
Introduction	6.16	9.27	8.19	8.33	7.33	7.05	7.23	8.20
Thesis	8.34	6.89	9.22	8.37	8.19	5.82	8.58	7.03
Main idea	20.52	11.92	16.48	6.95	15.11	7.13	17.37	9.14
Support	49.29	15.62	51.65	12.96	53.80	10.93	51.58	13.26
Conclusion	10.63	5.19	12.81	7.26	13.95	7.48	12.47	6.78
Total								
Introduction	5.23	7.35	8.65	9.23	7.22	7.21	7.03	8.07
Thesis	9.56	7.50	9.64	7.89	8.51	6.94	9.24	7.44
Main idea	18.73	12.05	16.35	9.77	17.67	9.63	17.58	10.54
Support	52.26	16.94	50.29	13.66	52.06	10.98	51.54	14.04
Conclusion	10.10	7.35	12.88	8.49	12.92	6.29	11.97	7.52

Table 28: Descriptive statistics for development by candidate group and test occasion

Finally, Table 29 shows that the autocorrelations for organisation and development are positive indicating that, generally, candidates who included particular discourse elements at each test occasion tended to include those elements at the following test occasion and vice versa. Additionally, those who devoted more words to any discourse element at any test occasion tended to devote more words to the same element in the following test occasion and vice versa.

	Occasions 1 and 2	Occasions 2 and 3
Organisation		
Introduction	.11	.09
Thesis	.14	.40**
Main idea	.49**	.70**
Supporting ideas	NA	NA
Conclusion	.22	.09
Development		
Introduction	.16	.07
Thesis	.26*	.06
Main idea	.15	.27*
Supporting ideas	.26*	.25*
Conclusion	.03	.21

* *p*<.05; ** *p*<.01; N=78

Table 29: Autocorrelations for discourse measures

3.2.7 Register

Table 30 reports descriptive statistics for each of the three register indices across test occasions and candidate groups. It shows that while there were no large differences across test occasions for any of the measures, there are some large differences across candidate groups. For example, candidates scoring 6 at test occasion 1 used fewer contractions (*M*= .06 contractions per 100 words) and more passive constructions (*M*= 1.22 per 100 words) than did those scoring 5 (*M*= .40 and *M*= .66) and 4 (*M*= .49 and *M*= .48, respectively). The patterns for nominalisation are less clear. In particular, candidates scoring 4 at test occasion 1 seem to have used more nominalisations at test occasion 2 (*M*= 3.42) and test occasion 3 (*M*= 3.22) than they did at test occasion 1 (*M*= 2.12). Candidates scoring 5, in contrast, used fewer nominalisations at test occasion 2 (*M*= 2.34) and test occasion 3 (*M*= 2.99) than they did at test occasion 1 (*M*= 4.36). Furthermore, candidates scoring 4 used nominalisations less frequently (*M*= 2.12) than did those scoring 5 (*M*= 4.36) and 6 (*M*= 4.18) at test occasion 1, but at test occasion 2, candidates scoring 5 used nominalisations less frequently (*M*= 2.34) than did those scoring 6 (*M*= 4.19). The patterns in Table 30 suggest that candidates scoring 6 at test occasion 1 tended to use features associated with informal speech style (i.e., contractions) less frequently and to use features associated with formal academic style (i.e., passive voice, nominalisation) more frequently than did candidates with lower initial writing scores. Furthermore, candidates scoring 4 at test occasion 1 increased the level of formality of their writing by using more nominalisations at test occasions 2 and 3 compared to test occasion 1, while candidates scoring 5 showed the opposite pattern in terms of this feature.

Candidate group	4		5		6		Total	
Test occasion	M	SD	M	SD	M	SD	M	SD
Occasion 1								
Contractions	0.58	0.67	0.47	0.84	0.06	0.20	0.37	0.66
Passivisation	0.40	0.63	0.42	0.45	1.22	0.83	0.68	0.75
Nominalisation	2.12	1.31	4.36	3.30	4.18	2.45	3.55	2.66
Occasion 2								
Contractions	0.48	0.69	0.29	0.64	0.07	0.17	0.28	0.57
Passivisation	0.37	0.50	0.81	0.83	1.19	0.68	0.79	0.75
Nominalisation	3.42	2.28	2.34	1.52	4.19	1.81	3.31	2.02
Occasion 3								
Contractions	0.40	0.55	0.44	0.74	0.04	0.11	0.29	0.56
Passivisation	0.68	0.70	0.75	0.77	1.25	0.79	0.89	0.79
Nominalisation	3.22	2.18	2.99	1.92	4.08	1.88	3.43	2.03
Total								
Contractions	0.49	0.64	0.40	0.74	0.06	0.16	0.31	0.60
Passivisation	0.48	0.62	0.66	0.71	1.22	0.76	0.79	0.77
Nominalisation	2.92	2.03	3.23	2.49	4.15	2.04	3.43	2.25

Table 30: Descriptive statistics for register measures by candidate group and test occasion

The autocorrelations (Pearson *r*) of the three register measures across test occasions are reported in Table 31. The table shows that the correlations are significant and positive for passive constructions and contractions indicating that, generally, candidates who used each of these two features frequently at each test occasion used it frequently at the following test occasion and vice versa. The correlations are weaker for nominalisations, perhaps because of the variation across candidate groups in the use of nominalisations across test occasions as noted above.

	Occasions 1 and 2	Occasions 2 and 3
Contractions	.24*	.51**
Passivisation	.34**	.45**
Nominalisations	.14	.14

* *p*<.05; ** *p*<.01; N=78

Table 31: Autocorrelations for register measures

Table 32 displays the MLM results for the three measures of register: contraction, passivisation, and nominalisation ratios. For contractions, Table 32 shows that two-thirds of the variance (.22 or 61%) was within candidates. The intercept variance (.14) was significant (X^2 = 225.09, *df*.= 77, *p*<.01), indicating that the ratio of contractions varied significantly across candidates. The results for Model 2 indicated that the ratio of contractions decreased by .04 contractions per 100 words, on average, on each succeeding test occasion, but this decrease was not statistically significant. Nor did the rate of change in contraction ratio vary significantly across candidates (X^2 = 74.19, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on contraction ratio at time 1. Consequently, the final model for contractions is Model 3.

As Table 32 shows, according to Model 3, the average contraction ratio for candidates with a writing score of 4 at test occasion 1 was .57 contractions per 100 words. There was a non-significant decrease in contraction ratio by .04 contractions per 100 words, on average, on each succeeding test occasion. Furthermore, candidate group (i.e., time 1 writing score) was significantly associated with contraction ratios. Specifically, for each increase of one band in initial writing scores, contraction ratios decreased by .21 contractions per 100 words. The rate of change in contraction ratios across test occasions did not very significantly across candidates (X^2 = 74.06, *df*.= 77, *p*>.05). Model 3 explained 25% of the between-person variance, but no within-person variance in contraction ratios.

For passivisation, Table 32 shows that two-thirds of the variance (.36 or 62%) was within candidates. The intercept variance (.22) was significant (X^2 = 218.68, df.= 77, p<.01), indicating that the ratio of passivisation varied significantly across candidates. The results for Model 2 indicated that the ratio of passivisation increased by .11 passive constructions per 100 words, on average, on each succeeding test occasion; this increase was statistically significant. However, the rate of change in passivisation ratio did not vary significantly across candidates (X^2 = 86.01, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on passivisation ratio at test occasion 1. Consequently, the final model for passivisation ratios included occasion at level 1 and candidate group at level 2.

As Table 32 shows, according to Model 3, the average ratio of passivisation for candidates with a writing score of 4 at test occasion 1 was .30 passive constructions per 100 words. There was a significant increase by .11 passive constructions per 100 words, on average, on each succeeding test occasion. Furthermore, candidate group was significantly associated with passivisation ratio. Thus, for each increase of one band in initial writing scores, the passivisation ratio increased by .38 passive constructions per 100 words. The rate of change in passivisation ratios over time did not very significantly across candidates (X^2 = 86.09, *df*.= 77, *p*>.05). Fit statistics indicated that Model 3 fits the data significantly better than Model 1 (X^2 = 23.45, *df*.= 2, *p*<.01). Model 3 explained 57% of the between-person variance, but only 6% of the within-person variance in passivisation ratio.

For nominalisation, Table 32 shows that the greatest majority of the variance (4.39 or 87%) was within candidates. The intercept variance (.64), though comparatively small, was significant (X^2 = 111.98, df.= 77, p<.01), indicating that the ratio of nominalisations varied significantly across candidates. The results for Model 2 indicated that the ratio of nominalisations decreased by .06 nominalisations per 100 words, on average, on each succeeding test occasion; but this decrease was not statistically significant. However, the rate of change in nominalisation ratio varied significantly across candidates (X^2 = 98.70, *df*.= 77, *p*<.05). Furthermore, Model 3 indicated that there was a significant effect of candidate group on nominalisation ratio at test occasion 1.

As Table 32 shows, the average ratio of nominalisation for candidates with a writing score of 4 at test occasion 1 was 2.93 nominalisations per 100 words. There was a non-significant decrease by .06 nominalisations per 100 words, on average, on each succeeding test occasion. However, the rate of change in nominalisation ratios across test occasions varied significantly across candidates (X^2 = 98.92, *df*.= 77, *p*<.05). Furthermore, candidate group was significantly associated with nominalisation ratio. Thus, for each increase of one band in initial writing scores, the nominalisation ratio increased by .56 nominalisations per 100 words. Fit statistics indicated that Model 3 fits the data significantly better than Model 1 (X^2 = 10.77, *df*.= 2, *p*<.01). Model 3 explained 18% of the between-person variance and 12% of the within-person variance in nominalisation ratio.

	Contractions			**Passivisation**			**Nominalisation**		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Mode1 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Fixed effects (SE)									
Intercept	.31** (.05)	.35** (.06)	.57** (.08)	.79** (.07)	.68** (.08)	.30** (.08)	3.43** (.16)	3.49** (.27)	2.93** (.29)
Candidate group			-.21** (.04)			.38** (.07)			.56** (.19)
Occasion		-.04 (.04)	-.04 (.04)		.11* (.05)	.11* (.05)		-.06 (.18)	-.06 (.18)
Random effects									
Between-candidate	.14	.16	.12	.22	.21	.09	.67	2.45	2.00
X^2 (df)	225.09** (77)	141.85** (77)	123.75** (76)	218.68** (77)	134.41** (77)	100.48** (76)	111.98** (77)	135.42** (77)	124.22** (76)
Occasion slope		.001	.001		.02	.02		.55	.56
X^2 (df)		74.19 (77)	74.06 (77)		86.01 (77)	86.09 (77)		98.70* (77)	98.92* (77)
Within-candidate	.22	.22	.22	.36	.34	.34	4.39	3.87	3.87
Model fit									
Deviance (#parameters)	392.70 (2)	398.01 (4)	387.93 (4)	509.92 (2)	510.38 (4)	486.46 (4)	1038.71 (2)	1036.71 (4)	1027.94 (7)
Model comparison: X^2 (df)		5.31 (2)	4.77 (2)		.46 (2)	23.45** (2)		2.01 (2)	10.77** (2)

* *p*<.05; ** *p*<.01; N=78

Table 32: MLM results for register measures

3.2.8 Interactional metadiscourse markers

Table 33 reports descriptive statistics for interactional metadiscourse markers, as well as their subcategories, across test occasions and candidate groups. Overall, there does not seem to be any difference across candidate groups or test occasions in terms of the ratio of interactional metadiscourse markers. However, three makers – hedges, self-mentions and boosters – seem to vary across candidate groups and test occasions. For example, candidates scoring 6 at test occasion 1 seem to have used more hedges and boosters on average (*M*= .36 hedges and .11 boosters per T-unit) than did those scoring 4 (*M*= .24 and .07) and 5 (*M*= .24 and .09, respectively). Furthermore, candidates scoring 5 at test occasion 1 seem to have used fewer self-mentions at test occasion 2 (*M*= .22 per T-unit) than they did at test occasion 1 (*M*= .40) and test occasion 3 (*M*= .38). They also seem to have used more self-mentions at test occasions 1 and 3 than did candidates scoring 6 (*M*= .20 and .19, respectively). Overall, it seems that candidates scoring 6 at test occasion 1, tended to use more hedges and boosters and fewer self-mentions than did candidates with lower initial writing scores.

Candidate group	4		5		6		Total	
Test occasion	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Occasion 1								
Interactional	1.17	0.62	1.12	0.69	1.07	0.50	1.12	0.60
Hedges	0.28	0.25	0.22	0.15	0.40	0.26	0.30	0.23
Boosters	0.08	0.07	0.11	0.11	0.13	0.13	0.11	0.10
Attitude markers	0.22	0.2	0.18	0.12	0.21	0.13	0.20	0.15
Self mention	0.38	0.27	0.40	0.35	0.20	0.21	0.33	0.30
Engagement markers	0.21	0.24	0.23	0.30	0.13	0.20	0.19	0.25
Occasion 2								
Interactional	1.03	0.76	0.86	0.50	1.14	0.74	1.01	0.68
Hedges	0.22	0.23	0.26	0.17	0.34	0.18	0.27	0.20
Boosters	0.08	0.12	0.07	0.06	0.13	0.12	0.09	0.10
Attitude markers	0.18	0.17	0.14	0.15	0.21	0.15	0.18	0.15
Self mention	0.34	0.26	0.22	0.24	0.28	0.34	0.28	0.28
Engagement markers	0.21	0.30	0.16	0.21	0.20	0.32	0.19	0.28
Occasion 3								
Interactional	0.85	0.63	1.08	0.65	0.87	0.37	0.93	0.57
Hedges	0.23	0.20	0.26	0.21	0.34	0.18	0.28	0.20
Boosters	0.06	0.06	0.08	0.09	0.08	0.08	0.07	0.07
Attitude markers	0.14	0.14	0.15	0.12	0.14	0.13	0.14	0.13
Self mention	0.27	0.26	0.38	0.34	0.19	0.14	0.28	0.27
Engagement markers	0.14	0.21	0.20	0.26	0.12	0.13	0.16	0.21
Total								
Interactional	1.01	0.68	1.02	0.62	1.03	0.56	1.02	0.62
Hedges	0.24	0.23	0.24	0.18	0.36	0.21	0.28	0.21
Boosters	0.07	0.09	0.09	0.09	0.11	0.11	0.09	0.10
Attitude markers	0.18	0.17	0.16	0.13	0.18	0.14	0.17	0.15
Self mention	0.33	0.27	0.33	0.32	0.22	0.25	0.30	0.28
Engagement markers	0.19	0.25	0.20	0.26	0.15	0.23	0.18	0.25

Table 33: Descriptive statistics for metadiscourse markers by candidate group and test occasion

The autocorrelations (Pearson *r*) of the interactional metadiscourse markers across test occasions are displayed in Table 34. Table 34 shows that the correlations are positive, except for attitude markers, indicating that, generally, candidates who used any of the markers frequently at each test occasion used these markers frequently at the following test occasion and vice versa.

	Occasions 1 and 2	Occasions 2 and 3
Interactional	.23*	.25*
Hedges	.31**	.51**
Boosters	.18	.24*
Attitude markers	.01	-.02
Self mention	.31**	.32**
Engagement markers	.14	.22

* *p*<.05; ** *p*<.01; N=78

Table 34: Autocorrelations for metadiscourse measures

Table 35 displays the MLM results for the ratios of interactional metadiscourse markers. The results for Model 1 indicated that about three quarters of the variance (.28 or 74%) was within candidates. The intercept variance (.10) was significant (X^2 = 162.74, df.= 77, p<.01), indicating that the ratio of interactional metadiscourse markers varied significantly across candidates. The results for Model 2 indicated that the ratio of interactional metadiscourse markers decreased by .09 markers per T-unit, on average, on each succeeding test occasion; this decrease was statistically significant. However, the rate of change in the ratio of interactional metadiscourse markers did not vary significantly across candidates (X^2 = 60.74, *df*.= 77, *p*>.05). Furthermore, Model 3 indicated that there was no significant effect of candidate group on the ratio of interactional metadiscourse markers at test occasion 1. Consequently, the final model for the ratios of interactional metadiscourse markers was Model 2. According to this model, the average ratio of interactional metadiscourse markers for all candidates at test occasion 1 was 1.11 markers per T-unit. There was a significant decrease in the ratio of markers by .09 marker per T-unit (or about 1 marker per 10 T-units), on average, on each succeeding test occasion. Model 2 explained 4% of the within-person variance, but no between-person variance in the ratio of interactional metadiscourse markers.

The MLM results for boosters, attitude markers and engagement markers were similar to those of all interactional metadiscourse markers, while those for hedges and self-mention were different. For hedges, Table 35 shows that three-quarters of the variance (.03 or 75%) was within candidates. The intercept variance (.01) was small but significant (X^2 = 191.52, df.= 77, p<.01), indicating that the ratio of hedges varied significantly across candidates. The results for Model 2 indicated that the ratio of hedges decreased by .01 hedges per T-unit, on average, on each succeeding test occasion, but this decrease was not statistically significant. However, the rate of change in the ratio of hedges varied significantly across candidates (X^2 = 130.35, *df*.= 77, *p*>.05). Model 3 indicated that there was a significant effect of candidate group on the ratio of hedges at test occasion 1. As Table 35 shows, the average ratio of hedges for all candidates scoring 4 at test occasion 1 was .23 hedges per T-unit. There was a non-significant decrease by .01 hedges per T-unit, on average, on each succeeding test occasion. However, the rate of change in the ratios of hedges varied significantly across candidates (X^2 = 130.36, *df*.= 77, *p*<.01). Furthermore, candidate group was significantly associated with the ratios of hedges at test occasion 1. For each increase of one band in initial writing scores, hedges increased by .06 hedges per T-unit. The final model explained 33% of the within-person variance and no between-person variance in the ratio of hedges.

	Interactional		Hedges			Self-mention		
	Model 1	**Model 2**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Fixed effects (SE)								
Intercept	1.02** (.05)	1.11** (.06)	.28** (.02)	.29** (.02)	.23** (.03)	.30** (.02)	.32** (.03)	.37** (.04)
Candidate group					.06** (.02)			-.05* (.02)
Occasion		-.09** (.04)		-.01 (.02)	-.01 (.02)		-.02 (.02)	-.02 (.02)
Random effects								
Between-candidate	.10	.12	.01	.03	.03	.03	.03	.03
X^2 (df)	162.74** (77)	107.81** (77)	191.52** (77)	193.89** (77)	183.27** (77)	192.45** (77)	130.87** (77)	124.25** (77)
Occasion slope		.0005		.008	.008		.0003	.0002
X^2 (df)		60.74 (77)		130.35** (77)	130.36** (77)		72.92 (77)	72.73 (77)
Within-candidate	.28	.27	.03	.02	.02	.05	.05	.05
Model fit								
Deviance (#parameters)	425.81 (2)	427.19 (4)	-81.75 (2)	-79.58 (4)	-83.03 (4)	52.98 (2)	59.12 (4)	59.52 (4)
Model comparison: X^2 (df.)		1.38 (2)		2.17 (2)	1.28 (2)		6.24* (2)	6.54* (2)

* p<.05; ** p<.01; N=78

Table 35: MLM results for metadiscourse markers

For self-mentions, Table 35 shows that most of the variance (.05 or 63%) was within candidates. The intercept variance (.03) was small but significant (X^2 = 192.45, df.= 77, p<.01), indicating that the ratio of self-mentions varied significantly across candidates. The results for Model 2 indicated that the ratio of self-mentions decreased by .02 self-mentions per T-unit, on average, on each succeeding test occasion; this decrease was not statistically significant, however. Nor did the rate of change in self-mention ratio vary significantly across candidates (X^2 = 72.92, *df*.= 77, *p*>.05). However, Model 3 indicated that there was a significant effect of candidate group on self-mention ratio at test occasion 1. Consequently, the final model for self-mention is Model 3. As Table 35 shows, the average ratio of self-mentions for candidates scoring 4 at test occasion 1 was .37 self-mentions per T-unit. There was a non-significant decrease by .02 self-mentions per T-unit, on average, on each succeeding test occasion. The rate of change in the ratios of self-mentions did not vary significantly across candidates (X^2 = 72.73, *df*.= 77, *p*>.05). However, candidate group was significantly associated with the ratio of self-mentions at test occasion 1. For each increase of 1 band in initial writing scores, self-mentions decreased by .05 self-mentions per T-unit. Fit statistics indicated that Model 3 fits the data significantly better than Model 1 (X^2 = 6.54, *df*.= 2, *p*<.05). Model 3 explained no within-person or between-person variance in the ratio of self-mentions, however.

3.3 Relationships between script linguistic characteristics and scores across test occasions

To address research question 4 concerning the relationships between the linguistic and discourse characteristics of repeaters' scripts, on the one hand, and their script scores, on the other, across test occasions, MLM was employed. Before conducting MLM analyses, however, two sets of correlational analyses were conducted. First, the correlations (Pearson *r*) between each linguistic measure and script scores for each test occasion were examined. Table 36 reports the results of these analyses.

To assess whether the strength of the association between a given linguistic feature and writing scores varied significantly across test occasions, the interactive calculator developed by Lee and Preacher (2013) to test the equality of two correlation coefficients obtained from the same sample was used. This calculator converts each correlation coefficient into a *z*-score using Fisher's *r*-to-*z* transformation and then compares the two estimates to find out if they differ significantly.

The following patterns emerge from the results in Table 36:

- ***Fluency*****:** Number of words per script was positively and significantly correlated with writing scores at test occasions 1 and 2 (*r*= .45 and .43, respectively). The correlation for test occasion 3 (*r*= .22) was non-significant and significantly weaker than those for test occasion 1 (*Z*= 2.11, *p*<.05) and test occasion 2 (*Z*= 2.18, *p*<.05). Overall, longer scripts tended to receive higher writing scores, particularly at test occasions 1 and 2.
- ***Accuracy*****:** The correlations between the ratio of errors and writing scores are negative and significant for all test occasions. The strength of the correlation between the ratio of all errors and writing scores did not vary significantly across test occasions. Overall, scripts with fewer errors tended to receive higher writing scores at each test occasion. This pattern is true for all errors and for each category of errors (i.e., grammar, usage, mechanics and style).
- ***Syntactic complexity*****:** Only NP density was significantly correlated with writing scores at each of the three test occasions, with the correlation on occasion 3 being higher. The correlations between writing scores, on the one hand, and left embeddedness and syntax similarity, on the other, were weak and non-significant. The strength of the correlation between each of the three syntactic complexity measures and writing scores did not vary significantly across test occasions. Generally, scripts with higher NP density indices tended to obtain higher scores.
- ***Lexical features*****:** All four lexical measures (lexical density, MTLD, AWL and word frequency) had significant correlations with writing scores at each of the three test occasions, except for lexical density at test occasion 2. Word frequency correlated negatively with writing scores, while the other three measures correlated positively with writing scores. The strength of the correlation between AWL and word frequency, on the one hand, and writing scores, on the other, varied significantly across test occasions. Specifically, the correlation between AWL and writing scores for test occasion 2 (*r*=.42) was significantly weaker than that for test occasions 1 (*r*=.64; *Z*= 2.90, *p*<.05) and 3 (*r*=.60; *Z*= -2.17, *p*<.05). Similarly, the correlation between word frequency and writing scores for test occasion 2 (*r*=-.45) was significantly weaker than that for test occasion 3 (*r*=-.65; *Z*= 2.35, *p*<.05). Overall, scripts that included more content words, more diverse words, longer words, and more low-frequency words tended to obtain higher writing scores.

	Test occasion 1	Test occasion 2	Test occasion 3
Fluency: Words per script	.45**	.43**	.22
Accuracy: All errors	-.54**	-.58**	-.63**
Grammar	-.43**	-.43**	-.54**
Usage	-.28*	-.32**	-.47**
Mechanics	-.43**	-.52**	-.46**
Style	-.36**	-.32**	-.45**
Complexity: Left embeddedness	.02	.22	.04
NP density	.27*	.29*	.46**
Syntax similarity	-.06	-.13	-.06
Lexis: Lexical density	.26*	.14	.31**
Lexical variation: MTLD	.44**	.43**	.54**
Lexical sophistication: AWL	.64**	.42**	.60**
Word Frequency	-.55**	-.45**	-.65**
Cohesion: All connectives	.01	-.15	-.08
Argument overlap	.01	-.19	-.26*
LSA overlap, sentences	-.04	.02	-.04
LSA overlap, paragraphs	.25*	.13	-.07
Discourse: Organisation			
Introduction	.29*	.09	.11
Thesis	-.12	-.11	.11
Main idea	.10	.12	.03
Conclusion	.25*	.25*	.04
Discourse: Development			
Introduction	.25*	.02	.11
Thesis	-.15	-.03	.06
Main idea	-.07	.16	-.20
Supporting ideas	-.02	-.12	.11
Conclusion	.18	.07	.15
Register: Contractions	-.32**	-.28*	-.29**
Passivisation	.45**	.43**	.29**
Nominalisations	.32**	.15	.24*
Metadiscourse: Interactional	-.07	.07	-.02
Hedges	.22	.24*	.23*
Boosters	.18	.19	.18
Attitude markers	-.04	.08	-.05
Self-mention	-.25*	-.07	-.15
Engagement markers	-.12	-.04	-.13

* *p*<.05; ** *p*<.01; N=78

Table 36: Correlations between linguistic features and writing scores by test occasion

- ***Coherence and cohesion*:** The correlations between writing scores and each of the four coherence and cohesion measures were weak for all test occasions, except for argument overlap for adjacent sentences which correlated negatively and significantly with writing scores at test occasion 3 and mean LSA overlap for adjacent paragraphs which correlated significantly and positively with writing scores at test occasion 1. The strength of the correlation between argument overlap and mean LSA overlap for adjacent paragraphs, on the one hand, and writing scores, on the other, varied significantly across test occasions. Specifically, the correlation between argument overlap and writing scores for test occasion 1 (*r*=.01) was significantly weaker than that for test occasion 3 (*r*=-.26; *Z*= 1.97, *p*<.05). By contrast, the correlation between mean LSA overlap for adjacent paragraphs and writing scores for test occasion 3 (*r*=-.07) was significantly weaker than that for test occasion 1 (*r*=.25; *Z*= 2.11, *p*<.05). Overall, it seems that scripts with lower argument overlap tended to obtain higher writing scores at test occasion 3, while scripts with higher mean LSA overlap for adjacent paragraphs tended to obtain higher writing scores at test occasion 1.

- ***Discourse structure*:** The presence and length of the introduction were significantly and positively correlated with writing scores at time 1 indicating that scripts which included an introduction that is relatively longer tended to receive higher scores than did those scripts with no introduction or a shorter one at time 1. Additionally, the presence of a conclusion correlated positively and significantly with writing scores at times 1 and 2 suggesting that scripts which included a conclusion tended to receive higher writing scores than those that did not include a conclusion at times 1 and 2. None of the other organisation and development measures correlated significantly with writing scores at any of the test occasions. The strength of the correlation between each of the organisation and development measures and writing scores did not vary significantly across test occasions.
- ***Register*:** The ratios of contractions and passivisation were significantly correlated with writing scores for all three test occasions. However, the correlations were positive for passivisation and negative for contractions. The nominalisation ratio was significantly correlated with writing scores at test occasions 1 and 3 only. However, the strength of the correlation between each of the three register measures and writing scores did not vary significantly across test occasions. Overall, scripts that included fewer contractions and more passive constructions and nominalisations tended to obtain higher writing scores at each test occasion.
- ***Interactional metadiscourse markers*:** The correlations between writing scores and the ratio of interactional metadiscourse markers were almost zero for all test occasions. All subcategories of metadiscourse markers correlated weakly with writing scores at all test occasions, except for hedges which correlated positively and significantly with writing scores at test occasions 2 and 3 and self-mention, which correlated negatively and significantly with writing scores at test occasion 1. Boosters also seem to correlate positively with writing scores, though the correlations were not significant. Scripts that included more hedges tended to receive higher scores at test occasions 2 and 3, while scripts that included more self-mention tended to receive lower scores at test occasion 1. However, the strength of the correlation between each of the metadiscourse measures and writing scores did not vary significantly across test occasions. Overall, the patterns of correlations in Table 34 suggest that scripts that included more hedges and boosters and fewer self-mentions tended to obtain higher writing scores than did the scripts that included fewer hedges and boosters and more self-mentions at each test occasion.

Second, the correlations (Pearson *r*) among all the linguistic measures in the study were examined for each test occasion. The results indicated the following:

- The correlations between the two measures of lexical sophistication, AWL and word frequency, was negative and high for all test occasions (range: -.83 to -.74) which, unsurprisingly, suggests that longer words were less frequent than shorter words.
- The correlations between two measures of coherence and cohesion, argument overlap for adjacent sentences and mean LSA overlap for adjacent sentences, were almost .70 for the three test occasions.
- The correlations among the remaining measures in the study were all below .60.

To reduce the number of variables to be included in MLM analyses, only those linguistic measures that have at least one significant correlation with writing scores on at least one test occasion were considered for inclusion. Additionally, only one of each pair of linguistic measures that were highly correlated (i.e., $r \geq .70$) was retained in MLM analyses. Thus, word frequency and mean LSA overlap for adjacent sentences, which correlated highly with AWL and argument overlap for adjacent sentences, respectively, were excluded. Consequently, the final set of variables that were selected for inclusion in MLM analyses to address research question 4 consisted of the following 13 linguistic features:

- *Fluency*: number of words per script
- *Accuracy*: ratio of all errors
- *Syntactic complexity*: NP density
- *Lexical features*: lexical density, MTLD and AWL
- *Coherence and cohesion*: argument overlap and LSA overlap for paragraphs
- *Register*: contractions, passivisation, and nominalisations
- *Interactional metadiscourse markers*: hedges and self-mention.

As noted earlier, in order to examine the relationships between the linguistic and discourse features of the scripts and writing scores across test occasions, several MLM models were estimated following Hox's (2002) recommendations. Table 35 displays the results for the various MLM models for writing scores. The result for Model 1 indicated that slightly less than half of the variance in writing scores (.44 or 46%) was within candidates. The intercept of 5.62 in Model 1 is simply the average writing score across all candidates and test occasions. The intercept variance (.52) was significant (X^2 = 352.93, df.= 77, p<.01), indicating that writing scores varied significantly across candidates. Model 2 added occasion as a linear predictor at level 1. The model predicts a value of 4.99 at test occasion 1 (i.e., average writing score across all candidates at test occasion 1), which increases by .63 band score, on average, on each succeeding test occasion. This increase is statistically significant. The occasion slope variance was small (.03), but significant (X^2 = 231.39, *df*.= 77, *p*<.01), indicating that the rate of change in writing scores over test occasions varied significantly across candidates.

	Model 1	Model 2	Final Model	
Fixed effects (SE)			**Unstandardised**	**Standardised**
Intercept	5.62** (.09)	4.99** (.09)	5.03** (.08)	
Occasion		.63** (.02)	.59** (.02)	.49
Fluency			.001** (.0004)	.07
NP density			.30* (.12)	.05
MTLD			.003** (.001)	.06
AWL			.27** (.08)	.09
Contractions			-.07* (.03)	-.04
Self-mention			.14** (.05)	.04
Random effects				
Between-candidate	.53	.67	.47	
X^2 (df= 77)	352.93**	2595**	1702.91**	
Occasion slope		.03	.02	
X^2 (df= 77)		231.39** (77)	172.99**	
Within-candidate	.44	.02	.03	
Model fit				
Deviance (#parameters)	592.57 (2)	233.32 (4)	234.69 (4)	
Model comparison: X^2 (df.)		359.24** (2)	357.87** (2)	

* *p*<.05; ** *p*<.01

Table 37: MLM results for writing scores

Next, several models were developed and evaluated to identify which among the 13 linguistic features listed above were significantly associated with writing scores across test occasions and which associations between linguistic features and scores varied significantly across candidates. The results of these models indicated that only five linguistic features had significant associations with writing scores across test occasions: number of words per script, NP density, MTLD, AWL, contraction ratio, and self-mention ratio. None of the associations between these linguistic features and writing scores varied significantly across candidates. Consequently, the final model included only occasion and these five features at level 1; no level-2 predictors were included in the model.

The final model specifies changes in writing scores as a function of test occasion and changes in the five linguistic characteristics of the scripts across test occasions. The last two columns of Table 37 display the results for the final model. The results show that the average writing score for all candidates at test occasion 1 is 5.03 and that there was a significant increase of writing scores by .59 bands, on average, on each succeeding test occasion. Number of words, NP density, MTLD, AWL, and self-mention ratio were all positively and significantly associated with writing scores, while contraction ratio was negatively and significantly associated with scores.

To allow comparisons of the coefficients of the five linguistic features, which were measured on different scales, the coefficients were standardised (following steps in Hox, 2002, p. 21). The standardised coefficients indicated that the change over occasions is the largest effect (.49). Among the five linguistic features, AWL has the highest effect (.09), followed by number of words (.07), MTLD (.06), NP density (.05), self-mention ratio (.04) and contractions (-.04).

Overall, longer scripts with higher AWL, higher MTLD index, greater NP density, more self-mentions, and fewer contractions tended to obtain higher writing scores. The strength of the relationships between each of the five linguistic features and writing scores did not vary significantly across candidates. Fit statistics indicated that the final model fits the data better than Model 1 (*X^2* = 357.87, *df*.= 2, *p*<.01). The final model explained 33% of within-person (i.e., across test occasions) variance and 30% of the between-person variance in writing scores. As the significant between-person variance (.47) and occasion slope variance (.02) indicate, much of the variance in writing scores between candidates and in the rate of change in writing scores over time across candidates is not explained by the final model. Other candidate factors and covariates may explain the remaining variance.

4 SUMMARY AND DISCUSSION

This study aimed to examine patterns of *change over time* in the linguistic and discourse characteristics of IELTS repeaters' responses to Writing Task 2. The study included 234 scripts written by a purposive sample of 78 candidates who differed in terms of their initial writing abilities and who each took IELTS Academic three times. Various computer programs were used to analyse the scripts in terms of various features related to candidates' grammatical (i.e., fluency, accuracy, syntactic complexity, and lexical features), discourse (i.e., coherence and cohesion, discourse structure), sociolinguistic (i.e., register), and strategic (i.e., interactional metadiscourse markers) choices. This section summarises and discusses the key findings in relation to each of the four research questions that guided the study.

4.1 Differences in the linguistic characteristics of scripts at bands 4, 5 and 6 at test occasion 1

Fluency

Scripts at different band levels at test occasion 1 varied significantly in terms of their length, with scripts scoring 4 being, on average, significantly shorter than those scoring 5 and 6. The difference between scripts scoring 5 and those scoring 6 at test occasion 1 was not significant. MLM results confirmed that candidate group had a significant effect on fluency at test occasion 1 such that for each one-band increase in writing scores at test occasion 1, there was a significant increase in script length by 38.23 words, on average. These findings are consistent with previous studies which found that high-scoring scripts tend to be significantly longer than low-scoring scripts (e.g., Cumming et al., 2005; Grant and Ginther, 2000; Frase et al., 1999; Mayor et al., 2007; Riazi and Knox, 2013). Generally, less proficient writers seem to produce shorter and less elaborated texts (cf. Hinkel, 2002). Similarly, in the context of L1 writing, Crossley et al. (2011) found that high-scoring scripts tend to be longer. Crossley et al. explained that this is not surprising since "longer texts afford writers the opportunity to elaborate sufficiently on topics and arguments in their essays and enhance central ideas, all characteristics of proficient writers" (p. 301).

Accuracy

As expected, scripts scoring 4 at test occasion 1 included significantly more errors per 100 words (for all error types) than did those scoring 5 and 6; the differences between scripts scoring 5 and 6 were not significant. MLM results confirmed that candidate group had a significant effect on accuracy at test occasion 1 such that for each one-band increase in writing scores at test occasion 1, there was a significant decrease in the number of errors by 5.36 errors (per 100 words), on average. Cumming et al. (2005) and Banerjee et al. (2007), also, found that candidates with higher writing scores, in the context of TOEFL and IELTS, respectively, tend to demonstrate greater linguistic accuracy.

Syntactic complexity

The scripts did not differ significantly in terms of left embeddedness (i.e., the mean number of words before the main verb of main clauses) and mean sentence syntactic similarity for all combinations across paragraphs, but higher-scoring scripts had significantly greater NP density. In particular, scripts at band score 6 had a significantly higher NP density indices than did those scoring 4 at test occasion 1. MLM results confirmed that candidate group had a significant effect on NP density at test occasion 1 such that for each increase of one band in writing scores at test occasion 1, NP density increased by .06. Two previous studies on IELTS reported similar findings. Riazi and Knox (2013) found that scripts with scores 5, 6 and 7 did not differ significantly in terms of syntactic complexity, measured in terms of left embeddedness. Banerjee et al. (2007) also did not find a significant association between syntactic complexity and the writing scores of scripts at IELTS band levels 3 to 8. Other studies, however, found that syntactic complexity was significantly associated with writing scores. Cumming et al. (2005), for example, found that TOEFL candidates with higher proficiency tended to write longer and more clauses, while Mayor et al. (2007) found that sentence complexity was one of the strongest predictors of high scores on IELTS writing tasks.

Similarly, in the context of L1 writing, Crossley et al. (2011) found that more advanced writers used more syntactically complex structures, as measured by NP density, compared to less proficient writers. As Banerjee et al. (2007) cautioned, syntactic complexity by itself may not be a good indicator of increased L2 proficiency as measured by IELTS. Additionally, the complexity measures used in this study may not be good indicators of increasing IELTS levels.

Lexical features

All four measures of lexical features (lexical density, MTLD, AWL and word frequency) varied significantly across band levels. The main significant differences concerned scripts scoring 6 compared to those scoring 4 and 5 at test occasion 1. Overall, scripts scoring 6, on average, included significantly more content words (i.e., higher lexical density), more diverse vocabulary (i.e., greater MLTD), longer words (i.e., higher AWL), and more low-frequency vocabulary than did scripts with lower writing scores (4 and 5) at test occasion 1. MLM results confirmed that candidate group had a significant effect on each of the four measures of lexical features. Thus, for each one-band increase in writing scores at test occasion 1, there was, on average, significant decrease in word frequency (by .10) and significant increases in lexical density (by .01), MTLD (by 10.94), and AWL (by .22 letters).

These findings are consistent with those reported in other studies on IELTS and TOEFL (e.g., Crossley et al., 2010; Cumming et al., 2005; Banerjee et al., 2007; Frase et al., 1999; Grant and Ginther, 2000; Riazi and Knox, 2013). In the context of IELTS, Riazi and Knox (2013) found that scripts with higher scores used significantly more low-frequency words and had greater lexical diversity (i.e., higher TTR), while Banerjee et al. (2007) found that scripts with higher scores had greater lexical density, variation (i.e., higher TTR), and sophistication (i.e., more low-frequency words) than did low-scoring scripts. As for TOEFL, Cumming et al. (2005) and Grant and Ginther (2000) found that candidates with higher writing scores used longer (i.e., higher AWL) and more varied words (i.e., higher TTR) than did low-scoring candidates. Finally, Crossley et al. (2010) found that lexical diversity (i.e., MTLD) and word frequency were significantly associated with overall ratings of L2 essay quality.

Coherence and cohesion

There were no significant differences between scripts with different band levels at test occasion 1 in terms of connectives density (i.e., number of connectives per 1000 words), coreference cohesion (i.e., argument overlap for adjacent sentences), and one measure of conceptual cohesion (i.e., mean LSA overlap for adjacent sentences). The other measure of conceptual cohesion – mean LSA overlap for adjacent paragraphs – varied significantly across band levels; scripts scoring 6 had a significantly higher index for mean LSAP overlap for adjacent paragraphs than did scripts scoring 4 at test occasion 1. MLM results confirmed these findings indicating that candidate group had a significant effect only on mean LSA overlap for adjacent paragraphs; for each one band increase in writing scores at test occasion 1, there was a significant increase in mean LSA overlap for adjacent paragraphs by .05, on average. Riazi and Knox (2013) also found no significant differences between IELTS scripts scoring 5, 6 and 7 in terms of connectives density and argument overlap. Similarly, in the context of L1 writing, McNamara et al. (2010) found that cohesion indices did not show significant differences between high- and low-scoring scripts.

Discourse structure

The main differences concerned the inclusion of an introduction and, to a lesser extent, a conclusion. A significantly higher proportion of the scripts scoring 6 included an introduction and a conclusion than did scripts scoring 5 and 4. Additionally, in terms of development, scripts with higher scores tended to include relatively longer introductions and conclusions than did scripts with lower scores at test occasion 1. The relative length of the other discourse elements did not vary significantly across band levels.

Register

Scripts scoring 6 included significantly fewer contractions and more passive constructions and nominalisations per 100 words than did scripts scoring 5 and 4 at test occasion 1. MLM results confirmed that candidate group had a significant effect on the ratios of contractions, passivisation and nominalisations. Specifically, for each increase of one band in writing scores at test occasion 1, contraction ratios decreased by .21 contractions per 100 words, while passivisation increased by .38 passive constructions and nominalisations increased by .56 nominalisations per 100 words. Overall, scripts scoring 6 tended to include significantly fewer features associated with informal speech style (i.e., contractions) and significantly more features associated with formal academic style (i.e., passive voice, nominalisation) than did scripts with lower writing scores at test occasion 1. Grant and Ginther (2000) also found that as proficiency level increases, candidates tended to use more passive constructions and nominalisations, suggesting that as L2 writers become more proficient in their L2, they develop a more sophisticated awareness of the genre of academic writing.

Interactional metadiscourse markers

There were no significant differences across scripts at different score levels in terms of the use of interactional metadiscourse markers. However, scripts scoring 6 included significantly more hedges and fewer self-mentions than did those scoring 5. There were no significant differences between scripts scoring 4 and those scoring 5 or 6 for any of the metadiscourse measures. MLM results indicated that though candidate group did not have a significant effect on the ratio of interactional metadiscourse markers, the ratios of hedges and self-mentions did vary significantly across candidate groups at test occasion 1. Specifically, for each increase of one band in writing scores at time 1, hedges increased by .06 hedges while self-mentions decreased by .05 self-mentions per T-unit.

Similarly, Grant and Ginther (2000) found that test-takers with higher proficiency tended to use more hedges and to qualify the claims that they are making in their texts more often than those with lower proficiency.

Overall, the findings of this study indicate that scripts with higher writing scores were more likely to have the following features than scripts with lower writing scores at test occasion 1:

- include an introduction and a conclusion
- be significantly longer
- have higher linguistic accuracy, syntactic complexity (as measured by NP density), lexical density, diversity and sophistication (i.e., more content words, higher MTLD and AWL and more low-frequency vocabulary), coherence and cohesion (as measured by mean LSAP overlap for adjacent paragraphs)
- include longer introductions and conclusions
- include fewer informal features (i.e., contractions) and more formal features (i.e., passive voice, nominalisation)
- have more hedges and fewer self-mentions.

Most of the significant differences in terms of the linguistic features examined in this study concerned scripts at band levels 4 and 6. Specifically, scripts scoring 4 were less likely to include an introduction or a conclusion and tended to be significantly shorter, to include significantly more errors per 100 words, to have significantly lower syntactic complexity in terms of NP density, lower lexical density, variation and sophistication indices, lower mean LSAP overlap for adjacent paragraphs, more contractions, and fewer passive constructions and nominalisations than did scripts scoring 6 at test occasion 1.

Additionally, scripts scoring 5 were less likely to include an introduction or a conclusion and tended to have significantly lower lexical density, variation and sophistication indices, shorter introductions and conclusions, more contractions, fewer passive constructions and nominalisation, fewer hedges, and more self-mentions than did scripts scoring 6 at test occasion 1. Scripts scoring 4 and 5 did not differ significantly in terms of the linguistic features examined in this study.

Overall, these findings suggest that markers are able to use the IELTS rating scale for Writing Task 2 to distinguish consistently several relevant writing aspects of candidates' scripts across band levels 4 and 5, on the one hand, and level 6, on the other.

4.2 Changes across test occasion in the linguistic characteristics of repeaters' scripts.

The autocorrelations for the different linguistic measures in the study across test occasions tended to be positive and high suggesting that the order of the candidates relative to each other in terms of most of these measures was somewhat stable across test occasions. However, some features exhibited some significant changes across test occasions, while others did not. Furthermore, the rate of change in some features across test occasion varied significantly across candidates, while the change rate for other features did not.

Fluency

MLM results indicated that there was a significant increase in script length (by 30.42 words on average) on each succeeding test occasion. The rate of change in fluency across test occasions varied significantly across candidates, however. For example, while candidates who produced longer scripts at each test occasion produced longer scripts at the following test occasion and vice versa, the differences between candidate groups in terms of script length decreased over time. The significant increase in script length across test occasions suggests that the candidates tended to elaborate their arguments and ideas more in subsequent test occasions. The finding that differences in fluency eventually attenuate is likely the result of the test having a time limit; even more proficient candidates can produce so many words within the time limits of the test.

Accuracy

MLM results indicated that there was a significant decrease in the number of errors (by 1.03 errors per 100 words, on average) on each succeeding test occasion. However, the rate of change in accuracy did not vary significantly across candidates and, generally, candidates who made more errors at each test occasion made more errors at the following test occasion and vice versa.

Syntactic complexity

MLM results indicated that none of the three measures of syntactic complexity (left embeddedness, syntax similarity and NP density) changed significantly across test occasions. Furthermore, only the rate of change for left embeddedness over time varied significantly across candidates. Autocorrelations indicated that the order of candidates relative to each other across test occasion was more stable for syntactic similarity than it was for left embeddedness and NP density.

Lexical features

MLM results indicated that the four lexical measures (lexical density, MTLD, AWL and word frequency) did not change significantly across test occasions. Furthermore, only the rate of change in lexical density over time varied significantly across candidates. Candidates with higher indices on each measure, particularly AWL and word frequency, at each test occasion had higher indices on that measure at the following test occasion and vice versa.

Coherence and cohesion

MLM results indicated that three of the four measures of coherence and cohesion (connectives density, argument overlap for adjacent sentences and mean LSA overlap for adjacent sentences) did not change significantly across test occasions. Mean LSA overlap for adjacent paragraphs, on the other hand, increased significantly (by .05, on average) over test occasions. Additionally, the rate of change in LSA overlap for adjacent sentences varied significantly across candidates. Finally, candidates who had higher indices on each of the four coherence and cohesion measures at each test occasion had higher indices at the following test occasion and vice versa.

Discourse structure

There were no significant changes in the proportion of candidates who included each of the five discourse elements in their scripts across test occasions. Nor did the relative length of the discourse elements vary significantly across test occasions. Generally, candidates who included particular discourse elements at each test occasion tended to include those elements at the following test occasion and vice versa. Additionally, candidates who devoted more words to any discourse element at any test occasion tended to devote more words to the same element in the following test occasion and vice versa.

Register

MLM results indicated that neither the contraction ratio, nor the nominalisation ratio changed significantly across test occasions. However, there was a significant increase by .11 passive constructions per 100 words, on average, on each subsequent test occasion. Furthermore, only the rate of change in nominalisation ratio varied significantly across candidates. The autocorrelations indicated that candidates who used passive constructions and contractions frequently at each test occasion used them frequently at the following test occasion and vice versa. The autocorrelations for nominalisations were weaker suggesting more variability in the use of this feature across candidate groups across test occasions.
For instance, candidates scoring 4 at test occasion 1 increased the level of formality of their writing by using more nominalisations at test occasions 2 and 3 compared to test occasion 1, while candidates scoring 5 at test occasion 1 showed the opposite pattern in terms of the use of this feature.

Interactional metadiscourse markers

MLM results indicated that there was a significant decrease in the ratio of all interactional metadiscourse markers by almost 1 marker per 10 T-units, on average, on each succeeding test occasion. The rate of change in the ratio of interactional metadiscourse markers did not vary significantly across candidates. However, some markers (e.g., hedges, self-mention) did not show a significant change across test occasions, but there was significant variability in terms of the rate of change in hedges across test occasions. Furthermore, candidates who used any of the markers frequently at each test occasion tended to use these markers frequently at the following test occasion and vice versa.

Overall, the findings of this study indicate that only six linguistic features changed significantly across test occasions. To answer the question raised in the title of this paper, it seems that the linguistic features that tended to change across test occasions are:

- script length
- the ratio of errors
- mean LSA overlap for adjacent paragraphs
- the number of passive constructions and, possibly, nominalisation
- the ratio of interactional metadiscourse markers.

Thus, scripts produced at subsequent test occasions tended to be significantly longer, more linguistically accurate (i.e., included fewer errors), more coherent (as measured by mean LSA overlap for adjacent paragraphs), and to include more formal features (i.e., more passive constructions and nominalisations) and fewer interactional metadiscourse markers than the scripts produced at earlier test occasions.

It should be noted here that because the test is timed, some differences across candidate groups might attenuate over time. For example, differences in fluency eventually attenuate because even more proficient candidates can produce so many words within the time limits of the test.

4.3 Effects of initial L2 writing ability on rate of change in the characteristics of repeaters' scripts

As noted above, MLM results indicated that the rate of change over time for six linguistic features varied significantly across candidates: fluency, left embeddedness, lexical density, mean LSA overlap for adjacent sentences, nominalisation ratio, and ratio of hedges. MLM analyses examined whether initial L2 writing ability (i.e., Writing Task 2 score at test occasion 1) can explain the variability across candidates in the rate of change over time in each of these features.

However, the results indicated that only the rate of change in mean LSA overlap for adjacent sentences was significantly moderated by initial L2 writing ability.

Specifically, the rate of change in mean LSA overlap for adjacent sentences was weaker (by .03) for each one-band increase in initial writing scores. This means that candidates with higher initial writing scores exhibited a lower rate of change in mean LSA overlap for adjacent paragraphs compared to candidates with lower initial writing scores.

4.4 Relationships between script linguistic characteristics and scores across test occasions

Research question 4 concerns the relationships between the linguistic and discourse characteristics of repeaters' scripts, on the one hand, and their Writing Task 2 scores, on the other, across test occasions. Correlational analyses indicated the following:

Fluency

Overall, longer scripts tended to receive higher writing scores, particularly at test occasions 1 and 2.

Accuracy

Scripts with fewer errors of all types tended to receive higher writing scores at each test occasion.

Syntactic complexity

Only NP density was significantly correlated with writing scores at each of the three test occasions, with the correlation on occasion 3 being higher. Generally, scripts with higher NP density indices tended to obtain higher scores.

Lexical features

Overall, scripts that included more content words, more diverse words, longer words, and more low-frequency words tended to obtain higher writing scores.

Coherence and cohesion

The correlations between writing scores and each of the four coherence and cohesion measures were weak for all test occasions.

Discourse structure

Scripts that included an introduction and a conclusion tended to receive higher scores.

Register

Scripts that included fewer contractions and more passive constructions and nominalisations tended to obtain higher writing scores at each test occasion.

Interactional metadiscourse markers

The correlations between writing scores and the ratio of interactional metadiscourse markers were almost zero for all test occasions. However, scripts that included more hedges and boosters and fewer self-mentions tended to obtain higher writing scores than did the scripts that included fewer hedges and boosters and more self-mentions at each test occasion.

In most cases, the strength of the correlation between the linguistic features and writing scores did not vary significantly across test occasions, except for fluency, AWL, word frequency, argument overlap, and mean LSA overlap for adjacent paragraphs. For fluency, the correlation for test occasion 3 was significantly weaker than those for test occasions 1 and 2, possibly because of the decrease in the differences in fluency between scripts at test occasion 3. For AWL, the correlation between AWL and writing scores for test occasion 2 was significantly weaker than those for test occasions 1 and 3. Similarly, the correlation between word frequency and writing scores for test occasion 2 was significantly weaker than that for test occasion 3. Finally, the correlation between argument overlap and writing scores for test occasion 1 was significantly weaker than that for test occasion 3, while the correlation between mean LSA overlap for adjacent paragraphs and writing scores for test occasion 3 was significantly weaker than that for test occasion 1.

MLM results indicated that, when all the features are considered together, only five linguistic features had significant associations with writing scores across test occasions:

- number of words per script
- NP density
- MTLD
- AWL
- contraction ratio
- self-mention ratio.

All features correlated positively with writing scores, except for contraction ratio which correlated negatively with scores. Among the five linguistic features, AWL has the highest effect (.09), followed by number of words (.07), MTLD (.06), NP density (.05), self-mention ratio (.04) and contraction ratio (-.04). Overall, longer scripts with higher lexical sophistication (i.e., AWL) and diversity (i.e., MTLD), higher syntactic complexity (as measured by NP density), more self-mentions, and fewer contractions tended to obtain higher writing scores.

Finally, while the correlations between some of the linguistic features and writing scores seem to vary across test occasions, reassuringly, the relationships between script linguistic features and scores did not seem to vary across candidates suggesting that the magnitude of the effects of the linguistic features included in the final model on writing scores was consistent across candidates.

These findings are consistent with previous studies on the linguistic characteristics of IELTS Writing Task 2 scripts (e.g., Banerjee et al., 2007; Mayor et al., 2007; Riazi and Knox, 2013). For example, Banerjee et al. (2007) and Riazi and Knox (2013) found that lexical diversity and lexical sophistication were significantly associated with scores on IELTS writing tasks, while Mayor et al. (2007)

found sentence complexity among the strongest predictors of IELTS writing scores (see Crossley et al., 2010 for similar findings in L2 writing, and Crossley et al., 2011 and McNamara et al., 2009 for similar findings in the context of L1 writing).

Like this study, Riazi and Knox (2013) found that cohesion indices were not significantly associated with writing scores. Similarly, in the context of L1 writing, McNamara et al. (2010) found that cohesion indices were not significantly correlated with writing scores. This does not mean that cohesions and coherence are not important characteristics of good writing. As McNamara et al. (2010) emphasised, we need to distinguish between cohesion, that is the cues that can be detected within the text, and coherence, which is in the mind of the reader. Thus, two texts may have low cohesion, but one is more coherent than the other. Additionally, a text may lack cohesion cues, but the reader or rater is still able to understand it by generating inferences to make sense of the text. In that sense, "the coherence of the [script] may emanate from some other aspects of the text that cannot be measured by [overt cohesion indices]" (p. 18). It is also possible that the measures of cohesion and coherence included in this study are not sensitive to differences in L2 proficiency (across candidates and test occasions) as measured by IELTS.

These observations have significant implications for automated approaches to writing assessment as well. Specifically, while only five linguistic features have been observed to correlate with writing scores in this study, it would be dangerous to assume that these five features in isolation could be used for marking writing as they fail to take into account the coherence (or lack thereof) of any piece of writing. Moreover, no one linguistic feature by itself is a good indicator of changes in L2 writing proficiency or overall judgment of writing quality.

As Connor-Linton and Polio (2014) cautioned, researchers examining change over time in the linguistic characteristics of L2 learners' scripts and the relationships between linguistic measures and overall ratings of L2 writing quality need to be aware that different measures may exhibit different patterns of change over time and display different patterns of relationships with writing scores (cf. Bulté and Housen, 2014; Crossley and McNamara, 2014; Polio and Shea, 2014). For instance, Crossley and McNamara (2014) found that while several linguistic measures showed significant changes over time in the scripts of a sample of L2 learners, these measures did not correlate significantly with human ratings, while Polio and Shea (2014) found that while language scores increased over time, there were no changes in the accuracy measures of the scripts of the same sample of L2 learners' over time.

5 LIMITATIONS

The findings above highlight several differences and changes in the linguistic characteristics of repeaters' scripts across candidate groups and test occasions. However, when interpreting these findings, several limitations of the study must be acknowledged; chief among them is the fact that this study is correlational. Most of these limitations also point to areas for further research.

First, the sample of candidates included in the study is neither large nor representative of the full range of candidates who usually take or repeat IELTS.

Second, the study included candidates with writing scores between 4 and 6 at test occasion 1. This could have affected the findings of the study (e.g., correlations among variables, rate of change in linguistic features across test occasions) by narrowing the range of writing proficiency levels included in the study. In particular, the restricted range of scores and proficiency levels included in the study could have attenuated the correlations among the variables in the study. Additionally, including a wider range of writing abilities (i.e., band levels 2 to 8, cf. Banerjee at el., 2007) could identify more and larger differences in the linguistic characteristics of scripts at different band scores and/or detect more variability in the rate and nature of changes in the linguistic characteristics of repeaters' scripts.

Third, the study included only three test occasions, which limited the range of analyses that could be conducted. For example, only linear change in linguistic features and writing scores were examined. Non-linear relationships could not be modelled. Furthermore, the effects of number of previous tests and the length of the interval between test occasions (i.e., practice effects) on changes in the linguistic characteristics of repeaters' scripts were not examined in this study. Future studies need to include a larger number of test occasions and candidates with a wider range of L2 and writing proficiency levels in order to better estimate changes in the linguistic characteristics of repeaters' scripts, as well as changes in the relationships between these characteristics and writing scores across test occasions.

Fourth, while the model adapted for analysing the scripts in this study was theoretically-sound and empirically-grounded, the study examined only those linguistic features that could be coded using the computer. Consequently, some important writing features (e.g., rhetorical structure, argument quality; cf. Cumming et al., 2005; Riazi and Knox, 2013) were not examined. Also, while computer programs can identify several key linguistic structures in candidates' scripts, they cannot evaluate whether these structures are used appropriately and accurately or not (Grant and Ginther, 2000).

This issue is further compounded by the fact that some features might not have been measured accurately by the computer programs used in this study. For example, recently, Lavolette, Polio and Kahng (2015) found that *Criterion* was able to identify only 54% of the errors manually identified in a sample of 128 essays by 32 L2 learners. Additionally, the errors that *Criterion* found were coded correctly 75% of the time; the remaining errors identified by *Criterion* were either correctly identified but miscoded (14%) or were for structures that were already correct (11%).

Similarly, estimates of the number of nominalisations per script computed by the computer program MAT could be unreliable, because MAT identifies all words with the specified endings listed in the program as instances of nominalisations even though some are not (e.g., comment, environment, nation, station). This could have affected the findings of the study. For example, the differences across groups and occasions in terms of nominalisations could have been over- or underestimated. That this study did not include other relevant linguistic and discourse features and that the measurement of some of the features may have lower accuracy could explain why a large proportion of the within- and between-person variance in writing scores was not explained by the features included in the study.

To address some of these limitations, the current set of 234 scripts could be further analysed qualitatively in terms of other linguistic and discourse features than those included in this study such as argument structure and quality, rhetorical structure, content, linguistic appropriacy, and coherence (cf. Cumming et al., 2005; Riazi and Knox, 2013). Automated analyses can reveal much, as this study has demonstrated, but they cannot, as yet, cope with the crucial qualities of argumentation and coherence in writing. Such analyses could focus on a smaller subset of the scripts, such as scripts by candidates who showed large score gains across test occasions and compare them to those who did not show any score gains.

Fifth, the study did not consider the effects of task and candidate factors on the linguistic characteristics of the scripts or the nature and rate of change in these features across test occasions. Obviously, each candidate responded to an equivalent, but not identical, writing task at each test occasion. Additionally, different candidates in the study responded to equivalent, but not identical, writing tasks, at the same test occasion. The differences between the different forms of the writing tasks administered to the candidates at different test occasions might have influenced the linguistic characteristics of their scripts. Task variability across candidates and test occasions could explain, for example, some of the remaining variance in some of the linguistic features (e.g., ratios of nominalisations, passivisation, self-mentions) between- and within candidates.

Nor were candidate variables, other than initial L2 writing ability (i.e., writing score at test occasion 1), considered in this study. Previous research has consistently shown that several of the linguistic characteristics of scripts vary significantly depending on candidate L1, overall English language proficiency, level of study (i.e., graduate or undergraduate), and context (e.g., ESL or EFL, country), to name a few factors. It is also possible that candidate factors (i.e., L1, age, gender, etc.) and context (ESL vs. EFL) affect not only the linguistic characteristics of repeaters' scripts but also the nature and rate of changes in these characteristics over time. Again, it is perhaps because the study did not include many candidate and task variables that most of the within- and between-person variance in the linguistic features examined in this study remained unexplained.

Consequently, it would be interesting to re-analyse the current dataset, or a small subset of it, to examine the relationships (if any) between changes in the linguistic characteristics of the scripts, on the one hand, and the characteristics of the writing tasks administered at each test occasion and candidate characteristics (e.g., L1, age), on the other. However, examining task and candidate effects requires including a larger and more varied sample of writing tasks and candidates from different contexts and backgrounds.

Sixth, the study did not collect data about what the candidates did between test occasions. While the interval between tests varied between 5 and 219 days, no information was available as to whether any of the candidates included in the study engaged in any activities to improve their English language proficiency and writing before or between tests. While one cannot expect that any serious language study and learning could have taken place during short intervals (i.e., less than a month), it is very likely that at least some of the participants in this study engaged in some activities to improve their English during long intervals between tests (the longest interval between any two tests was slightly more than five months). Furthermore, no information external to the test was available about the English language proficiency of the candidates, such as scores on another English language proficiency test or English course placement. Future studies could collect data on these variables (e.g., what language study activities individuals undertake between test occasions) and examine their relation to changes in writing performance across test occasions.

Finally, because the study included the final score for each script, it was not possible to examine whether the relationships between script linguistic features and scores varied across raters and across occasions for the same rater. Future studies could examine the relationships between changes in the linguistic characteristics of repeaters' scripts and the original ratings assigned by individual raters across test occasions, as well as whether and how the rating criteria and processes that raters employ vary over time (cf. Barkaoui, 2010a).

Despite its limitations, the study provides detailed, empirically-based descriptions of the writing features that distinguish scripts at different IELTS proficiency levels. This information is consistent with findings from previous studies and confirms that markers are able to use the IELTS rating scale for Writing Task 2 to distinguish consistently several relevant writing aspects of candidates' scripts across band levels 4 and 5, on the one hand, and band level 6, on the other. However, it seems that it is the grammatical (i.e., fluency, accuracy, lexical features) and sociolinguistic features (i.e., contractions, passivisation, nominalisation) that were more sensitive to differences across band levels than were the discourse (i.e., cohesion and coherence, discourse structure), syntactic complexity, and strategic features (i.e., metadiscourse use) (cf. Riazi and Knox 2013).

Additionally, the analyses did not detect any significant differences across band levels 4 and 5 in terms of the linguistic features included in this study, either because scripts at levels 4 and 5 do not differ significantly in terms of their linguistic characteristics or because the measures used in this study were not sensitive to differences across these two levels as defined by the rating scale for IELTS Writing Task 2. This is an area for further research in order to better define the features that distinguish scripts scored at levels 4 and 5 and better understand the role of discourse features in distinguishing scripts at different band levels. The findings of this study also identify several of the language and discourse aspects and abilities that are engaged by IELTS Writing Task 2 (cf. Banerjee et al., 2007).

A key contribution of the current study is that it provides an initial description of the patterns of change in IELTS repeaters' scripts across test occasions and how these changes relate to candidate initial L2 writing proficiency. Previous studies have examined only changes in repeaters' test scores (e.g., Green, 2005). The findings of this study indicated that some features of repeaters' scripts (e.g., fluency, linguistic accuracy) do change across test occasions, while other features (e.g., cohesion) do not. MLM results showed that writing scores did increase significantly (by .63 band level, on average) across test occasions, indicating that the overall quality of the scripts of the sample included in the study did improve over time. Nevertheless, only a handful of the linguistic features examined in this study exhibited significant changes across test occasions.

One possible explanation is that some linguistic features that exhibited change do not relate to scores significantly. For example, while mean LSA overlap for adjacent paragraphs showed significant change across test occasions, this feature did not correlate significantly with writing scores. By contrast, script length showed significant change across test occasions and was significantly associated with writing scores, although the association became weaker in test occasion 3. This suggests that some script features may contribute more to score variance at earlier stages of development than at later stages. It is also possible that some other features exhibit the opposite pattern; that is, they become more important in explaining score variance at later stages of development. Another issue is that overall writing scores and analyses of individual linguistic features might be sensitive to different aspects of change in writing performance. For example, analyses of individual linguistic features can detect fine-grained changes (and differences) in specific features (e.g., use of passive voice) that overall scores do not detect or reflect, while overall scores can detect changes (and differences) that fine-grained analyses of individual linguistic features cannot detect, such as the simultaneous change in multiple linguistic features that, when considered separately, do not show significant improvement, but, when considered together, as when reading and evaluating a piece of writing, can enhance the overall quality of the script as perceived by the reader or rater significantly. This issue applies to cross-sectional studies comparing the sensitivity of writing tests to differences across candidates at one point in time as well. In these studies, too, variance in individual linguistic features may not explain differences in overall writing scores.

Finally, the linguistic features that did not show change may require more time (and instruction) to develop. Consequently, candidates (and their instructors) need to be aware that some writing aspects take longer to develop and that candidates need to take this into account before attempting the test again.

6 IMPLICATIONS FOR FUTURE RESEARCH

The points raised above have several implications for future research on IELTS Writing Task 2.

6.1 Detecting true changes in the linguistic features of responses

First, there is a need for more research on whether and how the rating scales and rating procedures of IELTS Writing Task 2 can detect true changes in the linguistic features of candidates' responses across test occasions; which aspects of the rating scale are sensitive to change in writing ability; and whether raters are able to detect changes in writing ability over time. This research needs to identify and better operationalise the key writing features included in the rating scale for Writing Task 2 and how raters interpret and use them.

Future studies need also to examine whether the relative importance of different criteria on the rating scale changes over time and whether such changes are due to true changes in the combined effects of multiple features on the overall quality of the script or to other (construct-irrelevant) factors such as changes in task characteristics and requirements or changes in raters' interpretations of the rating criteria and reactions to writing tasks and candidate characteristics (e.g., L1). For example, raters may become used to the writing of candidates from particular backgrounds and, as a result, they become more or less severe when marking scripts written by candidates from that background. In this case, changes in writing scores do not reflect true changes in candidate writing ability as much as changes in raters' perceptions. This line of research can inform revisions of the rating scale, rating procedures and rater training for Writing Task 2 in order to improve the test's sensitivity to both differences across candidates and changes over time in candidate writing ability.

6.2 Examining changes before and after language instruction

Furthermore, future studies need to examine changes in the linguistic characteristics of scripts written by L2 learners before and after English language instruction. As noted above, previous studies on IELTS looked only at score gains after instruction. Research on changes in the linguistic characteristics of candidates' scripts in relation to L2 instruction can help assess whether and to what extent Writing Task 2 is sensitive to changes over time in writing proficiency and to the effects of different contexts (e.g., ESL vs. EFL) and types of writing instruction on writing proficiency. Such research needs to combine linguistic analyses with qualitative methods (e.g., interviews, observation) in order to find out, not only what writing aspects change and which do not, but also why and how students' writing develops over time and in relation to L2 instruction. While this is not the intended purpose of the test, some programs may use IELTS to measure writing development or progress over time and/or in relation to instruction.

In order to be able to use the test to make valid claims about L2 development, more research needs to be conducted that examines the test's sensitivity to change and instruction effects. Because sensitivity to change is a function of task type and requirements, rating scale criteria and levels, and rater training and rating procedures, all these aspects of the test need to be investigated longitudinally.

6.3 Implications for test validation and SLA research

Finally, the study has implications for test validation and second language acquisition (SLA) research.

First, the study has demonstrated how repeaters' test performance can be examined by combining text and score analyses. Specifically, the various measures used in this study allowed the detection of differences across band levels and test occasions in terms of specific linguistic features as well as the examination of changes in the linguistic characteristics of repeaters' scripts and how these changes relate to changes in repeaters' writing scores across test occasions. In doing so, the study has demonstrated the value and process of examining test repeater data as part of a larger program of test validation. Theoretically, the findings of the study describe some of the linguistic features that distinguish writing performance at various levels of achievement in L2 learning and identify some of the linguistic features of L2 learners' texts that change over time and how differences in initial L2 writing ability relate to changes in the linguistic characteristics of L2 learners' texts. These analyses were supported by Connor and Mbaye's (2002) framework which provides a theoretically-sound and empirically-grounded conceptualisation of writing ability as consisting of four main components.

This framework was operationalised in this study to examine variability in the linguistic characteristics of repeaters' scripts across candidates and test occasions, but it could be used in future studies to examine variability in L2 learners' texts across tasks, contexts, individuals, and time (cf. Barkaoui and Knouzi, 2012). For example, future studies using this framework could adopt a longitudinal design to examine the relationships between: (a) amount and nature of English language instruction; (b) changes in learner English language proficiency; (c) changes in the linguistic characteristics of L2 learners' texts; and (d) changes in their writing scores over time. Such a program of research needs to include more than three measurement occasions and large samples of candidates.

To date, only a handful of studies have examined the linguistic characteristics of L2 learners' texts before and after L2 instruction (e.g., Bulté and Housen, 2014; Crossley and McNamara, 2014; Friginal and Weigle, 2014; Polio and Shea, 2014; Storch, 2009).

It is hoped that a longitudinal approach to examining performance on L2 writing tests in relation to L2 instruction could significantly strengthen the connections between learning, teaching and assessment and enhance the cross-fertilisation of theories and methods in language testing and SLA research.

REFERENCES

Anthony, L. (2012). *AntConc* Version 3.3.5 (computer software), Tokyo, Japan, Waseda University. Available from: http://www.laurenceanthony.net/

Anthony, L. (2013). Developing AntConc for a new generation of corpus linguists, *Proceedings of the Corpus Linguistics Conference CL 2013,* Lancaster University, UK pp. 14–16, Available at: http://www.laurenceanthony.net/research/20130722_26_cl_2013/cl_2013_paper_final.pdf

Anthony, L. and Bowen, M. (2013). The language of mathematics: A corpus-based analysis of research article writing in a neglected field, *Asian ESP Journal, 9*2, pp. 5–25

Banerjee, J., Florencia, F. and Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels, *IELTS Research Reports vol. 7*, pp. 241–309. Canberra: IELTS Australia and London: British Council

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence, *TESOL Quarterly*, *vol. 26*, pp. 390–395

Barkaoui, K. (2007). Participants, texts, and processes in second language writing assessment: A narrative review of the literature, *The Canadian Modern Language Review, vol. 64*, pp. 97–132

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods cross-sectional study, *TESOL Quarterly, vol. 44*, pp. 31–57

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modelling approach, *Language Testing, vol. 27*, pp. 515–535

Barkaoui, K. (2013). An introduction to multilevel modelling in language assessment research, *Language Assessment Quarterly, vol.* 10, pp. 241–273

Barkaoui, K. (2014). Quantitative approaches to analyzing longitudinal data in second-language research, *Annual Review of Applied Linguistics, vol. 34*, pp. 65–101

Barkaoui, K. and Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in L2 writing assessment. In E. Van Steendam, M. Tillema, G. Rijlaarsdam and H. van den Bergh (Eds), *Measuring writing: Recent insights into theory, methodology and practices,* vol 23 of *Studies in Writing*. Amsterdam: Elsevier

Bell, H. (2003). *Using frequency lists to assess L2 texts*, Unpublished PhD thesis, University of Wales, Swansea

Biber, D. (1988). *Variation across speech and writing*, Cambridge: Cambridge University Press

Brown, J. D. (1998). An investigation into approaches to IELTS preparation, with particular focus on the academic writing component of the test, *IELTS Research Reports, vol. 1*, pp. 20–37. Canberra: IELTS Australia and London: British Council

Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity, *Journal of Second Language Writing, vol. 26*, pp. 42–65

Canale, M. (1983). From communicative competence to communicative performance. In J. Richards and R. Schmidt (Eds), *Language and communication*. New York: Longman

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics, vol. 1*, pp. 1–47

Chang, Y-Y. and Swales, J. (1999). Informal elements in English academic writing: Threats or opportunities for advanced non-native speakers? In C. N. Candlin and K. Hyland (Eds), *Writing: Texts, processes and practices* pp. 145–167. New York: Routledge

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (Eds), *Building a validty argument for the Test of English as a Foreign Language,* pp. 319–352, New York: Routledge

Connor, U. and Mbaye, A. (2002). Discourse approaches to writing assessment, *Annual Review of Applied Linguistics*, *vol.* 22, pp. 263–278

Connor-Linton J. and Polio C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus, *Journal of Second Language Writing, vol. 26*, pp. 1–9

Crismore, A. Markkanen, R. and Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students, *Written Communication, vol. 10*, pp. 39–71

Crossley, S. A., Greenfield, J. and McNamara, D. S. (2008). Assessing text readability using cognitively based indices, *TESOL Quarterly*, vol *42*, pp. 475–493

Crossley, S. A. and McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification, *Journal of Second Language Writing,* vol *20*, pp. 271–285

Crossley, S. A. and McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners, *Journal of Second Language Writing,* vol *26*, pp. 66–79

Crossley, S. A., Louwerse, M., McCarthy, P. M. and McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts, *Modern Language Journal, vol 91*, pp. 15–30

Crossley, S. A., Salsbury, T. and McNamara, D. S. (2009). Measuring L2 lexical proficiency using hypernymic relationships, *Language Learning, vol 59*, pp. 307–334

Crossley, S. A., Salsbury, T., McNamara, D. S. and Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices, *Language Testing, vol 28*, pp. 561–580

Crossley, S. A., Salsbury, T., McCarthy, P. M. and McNamara, D. S. (2008). Using latent semantic analysis to explore second language lexical development. In D. Wilson and G. Sutcliffe (Eds), *Proceedings of the 21st international Florida artificial intelligence research society,* pp. 136–141. Menlo Park, California: AAAI Press

Crossley, S. Weston, J., McLain Sullivan, S. and McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis, *Written Communication, vol 28*, pp. 282–311

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K. and James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL, *Assessing Writing, vol 10*, pp. 5–43

Elder, C. and O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gain on IELTS, *IELTS Research Reports*, vol 4, Ed. R. Tulloh, pp. 207–254. Canberra: IELTS Australia Pty Limited

Ellis, N. (2002). Frequency effects in language processing, *Studies in Second Language Acquisition, vol 24*, pp. 143–188

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions, *Journal of Second Language Writing, vol 4*, pp. 139–155

Field, A. (2009). *Discovering statistics using SPSS,* 3rd ed, Thousand Oaks, CA: SAGE, Publications

Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis, *Discourse Processes, vol 25,* pp. 285–307

Frase, L. T., Faletti, J., Ginther, A. and Grant, L. (1999). Computer analysis of the TOEFL test of written English, *TOEFL Research Report N 64*, Princeton, New Jersey: Educational Testing Service

Friginal, E. and Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis, *Journal of Second Language Writing*, *vol 26*, pp. 80–95

Graesser, A. C., McNamara, D. S., Louwerse, M. and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language, *Behavior Research Methods, Instruments, and Computers, vol 36*, pp. 193–202

Grant, L. and Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences, *Journal of Second Language Writing*, *vol 9*, pp. 123–145

Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test, *Assessing Writing, vol 10*, pp. 44–60

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. London: Longman

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts, *TESOL Quarterly, vol 37*, pp. 275–301

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, vol 4*, pp. 195–202

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. New York: Continuum

Hyland, K. and Tse, P. (2004). Metadiscourse in academic writing: A reappraisal, *Applied Linguistics, vol 25*, pp. 156–177

IELTS. (2009). IELTS Candidate Performance 2009, *Research Notes, vol 40*, pp. 26–29

Intaraprawat, P. and Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays, *Journal of Second Language Writing, vol 4*, pp. 253–272

Kennedy, C. and Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS academic writing task. In L. Taylor and P. Falvey (Eds), *IELTS collected papers: Research in speaking and writing assessment,* pp. 316–376. Cambridge: Cambridge University Press

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction, vol 1*, pp. 60–69

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance, *Journal of Second Language Writing, vol 20*, pp. 148–161

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review, vol 104*, pp. 211–240

Landauer, T. K., Foltz, P. W. and Laham, D. (1998). Introduction to latent semantic analysis, *Discourse Processes, vol 25*, pp. 259–284

Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. (Eds). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production, *Applied Linguistics, vol 16*, pp. 307–322

Lavolette, E., Polio, C. and Kahng, J. M. (2015). The accuracy of computer-assisted feedback and students' responses to it, *Language Learning & Technology, vol 19*

Lee, I. A. and Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with no variable in common* (computer software), Available from http://quantpsy.org

Lim, H. and Kahng, J. (2012). Review of Criterion, *Language Learning and Technology, 16*2, pp. 38–45, Available from: http://llt.msu.edu/issues/june2012/review4.pdf

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition, *International Journal of Corpus Linguistics, vol 14*, pp. 3–28

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing, *International Journal of Corpus Linguistics, vol 15*, pp. 474–496

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development, *TESOL Quarterly, vol 45*, pp. 36–62

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives, *The Modern Language Journal, vol 96*, pp. 190–208

Luke, D. A. (2008). Multilevel growth curve analysis for quantitative outcomes. In S. Menard (Ed), *Handbook of longitudinal research: Design, measurement, and analysis,* pp. 545–564. New York: Academic Press

Malvern, D. and Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity, *Language Testing, vol 19*, pp. 85–104

Mayor, B., Hewings, A., North, S., Swann, J. and Coffin, C. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS academic writing task 2. In L. Taylor and P. Falvey (Eds), *IELTS collected papers: Research in speaking and writing assessment,* pp. 250–313. Cambridge: Cambridge University Press

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, *Behavioral Research Methods, vol 42*, pp. 381–392

McNamara, D. S., Cai, Z. and Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch (Eds), *Handbook of latent semantic analysis,* pp. 379–400. Mahwah, NJ: Erlbaum

McNamara, D. S., Crossley, S. A. and McCarthy, P. M. (2010). Linguistic features of writing quality, *Written Communication, vol 27*, pp. 57–86

Meara, P. and Bell, H. (2001). P_Lex A simple and effective way of describing the lexical characteristics of short L2 texts, *Prospect*, *16,* pp. 5–24

Nini, A. (2014). *Multidimensional Analysis Tagger 1,2 – Manual*. Available from: http://sites.google.com/site/multidimensionaltagger

O'Laughlin, K. and Arkoudis, S. (2009). Investigating IELTS exit score gains in higher education, *IELTS Research Reports, vol 10*, Ed. J. Osborne pp. 1–86. Canberra: IELTS Australia and London: British Council

Polio, C. (1997). Measures of linguistic accuracy in second language writing research, *Language Learning, vol 47,* pp. 101–143

Polio, C. (2001). Research methodology in second language writing research: The case text-based studies. In T. Silva and P. K. Matsuda (Eds), *On second language writing,* pp. 91–115. Mahwah, NJ: Lawrence Erlbaum

Polio, C, and Shea, MC, 2014, An investigation into current measures of linguistic accuracy in second language writing research, *Journal of Second Language Writing,* vol *26*, 10-27

Preacher, K. J., Wichman, A. L., MacCallum, R. C. and Briggs, N. E. (2008). *Latent growth curve modeling*. Los Angeles, CA: Sage

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T. and Bridgeman, B. (2012). Evaluation of the e-rater scoring engine for the TOEFL independent an integrated prompts, *ETS Research Report 12-06.* Princeton, NJ: Educational Testing Service

Rao, C., McPherson, K., Chand, R. and Khan, V. (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the GT reading and writing test module, *IELTS Research Reports vol 5*, pp. 237–262. Canberra: IELTS Australia and London: British Council

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. and Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modelling*. Lincolnwood, IL: Scientific Software International

Read, J. and Hayes, B. (2003). The impact of the IELTS test on preparation for academic study in New Zealand, *IELTS Research Reports vol 4,* Ed. R. Tulloh, pp. 153–206. Canberra: IELTS Australia Pty Limited

Read, J. (2005). Applying lexical statistics to the IELTS speaking test, *Research Notes, 20*, pp. 10–16

Riazi, A. M. and Knox, J. S. (2013). An investigation of the relations between candidates' first language and the discourse of written performance on the IELTS Academic Writing Test, Task 2, *IELTS Research Reports vol 2*, pp. 1–89. Canberra: IELTS Australia and London: British Council

Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth, *Applied Linguistics vol 26*, pp. 317–342

Shaw, P. and Liu, E. (1998). What develops in the development of second-language writing? *Applied Linguistics, vol 19*, pp. 225–254

Singer, J. D. and Willett, J. B. (2003). *Applied longitudinal data analysis: Modelling change and event occurrence*. Oxford: Oxford University Press

Storch, N. (2009). The impact of studying in a second language L2 medium university on the development of L2 writing, *Journal of Second Language Writing, vol 18*, pp. 103–118

Taylor, L. (2004). Second language writing assessment: Cambridge ESOL's ongoing research agenda, *Research Notes, vol 16,* pp. 2–3

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability, *Language Testing, vol 27*, pp. 335–353

Weigle, S. C. *(*2011). Validation of automated scores of TOEFL iBT tasks against nontest indicators of writing ability*, TOEFL iBT Research Report 15*. Princeton, NJ: Educational Testing Service

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

Wolfe-Quintero, K., Inagaki, S. and Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii