5. Documenting features of written language production typical at different IELTS band score levels

Authors

Jayanti Banerjee
Lancaster University

Florencia Franceschina
Lancaster University

Anne Margaret Smith
Lancaster University

CONTENTS

ABSTRACT

Grant awarded Round 10, 2004

This study addresses the question of how competence levels, as operationalised in a rating scale, might be related to what is known about L2 developmental stages.

This study has taken its lead from discussions about the benefit of collaboration between researchers in language testing and second language acquisition (eg Bachman and Cohen, 1998; Ellis, 2001; and Laufer, 2001). It addresses the question of how competence levels, as operationalised in a rating scale, might be related to what is known about L2 developmental stages. Looking specifically at the writing performances generated by Tasks 1 and 2 of the IELTS Academic Writing module, the study explores the defining characteristics of written language performance at IELTS bands 3–8 with regards to: cohesive devices used; vocabulary richness; syntactic complexity; and grammatical accuracy. It also considers the effects of L1 and writing task type on the measures of proficiency explored.

The writing performances of 275 test-takers from two L1 groups (Chinese and Spanish) were transcribed and then subjected to manual annotation for each of the measures selected. Where automatic or semi-automated tools were available for analysis (particularly in the area of vocabulary richness), these were used. The results suggest all except the syntactic complexity measures investigated here are informative of increasing proficiency level. Vocabulary and grammatical accuracy measures appear to complement each other in interesting ways. L1 and writing tasks seem to have critical effects on some of the measures, so they are an important factor to take into account in further research.

IELTS RESEARCH REPORTS, VOLUME 7, 2007

Published by © British Council 2007 and © IELTS Australia Pty Limited 2007

The research and opinions expressed in this volume are of individual researchers and do not represent the views of IELTS Australia Pty Limited or British Council. The publishers do not accept responsibility for any of the claims made in the research.

National Library of Australia, cataloguing-in-publication data, 2007 edition, IELTS Research Reports 2007 Volume 7
ISBN 978-0-9775875-2-0 Copyright 2007

AUTHOR BIODATA

JAYANTI BANERJEE

Jayanti Banerjee is a lecturer at Lancaster University. She has published in *Language Teaching* and the *Journal of English for Academic Purposes*. She has also contributed chapters to edited collections such as *Experimenting with uncertainty: Essays in honour of Alan Davies*, C Elder et al (eds) (2001), Cambridge University Press. Her main interests are language testing and assessment and English for academic purposes.

FLORENCIA FRANCESCHINA

Florencia Franceschina is a lecturer at Lancaster University. Her main research interests are language acquisition and its relation to theoretical linguistics, especially learnability, route of development and the influence of a speaker's first language on second language development and attainment in the domain of morphosyntax. Her empirical work has focused on the acquisition of syntax in very advanced second language speakers, and this is the theme of her monograph *Fossilized second language grammars* (2005, John Benjamins). She and other colleagues (including the co-authors of this report) are currently developing a longitudinal corpus of L2 writing.

ANNE MARGARET SMITH

Anne Margaret Smith has completed her PhD, which was supervised jointly by the departments of Linguistics and English Language and Educational Research at Lancaster University. Her thesis is a synthesis of two of her main research interests – teacher training (for language teachers) and inclusive education. Other interests include second language acquisition, learning differences and disabilities, and teacher expertise.

1 INTRODUCTION

1.1 Context

The fields of language testing and second language acquisition (SLA) have regularly and publicly discussed the benefits of co-operation (cf Hyltenstam and Pienemann, 1985; Bachman and Cohen, 1998; Shohamy, 1998; Ellis, 2001; Douglas, 2001 and Laufer, 2001). One area that stands to benefit from collaborative research is the development of performance scales and rating scales. In particular, we might ask how the operationalisation of competence levels, as expressed in a rating scale, is related to what is known about L2 developmental stages and what the profile of linguistic proficiency might be of students who perform at different levels of the scale.

1.2 Research rationale

The International English Language Testing System (IELTS) Writing scales have recently been revised towards a more analytical style (Shaw, 2002, pp 12). Consequently, the availability of more detailed descriptions of written language ability at each band level seems highly desirable. In a report of revisions to the IELTS Writing assessment criteria and scales, Shaw (2004) lists some of the key features that a good scale should have.

Among the desiderata is a scale's ability to:

- capture the essential qualities of learner written performance
- accurately describe how writing abilities progress with increasing proficiency
- clearly distinguish all the band levels.

Clearly, the better our understanding of what these essential qualities are, how they are manifested at different levels, and how sensitive they are to performance factors such as task effect, the better we will understand the L2 writing construct (eg Weigle, 2002; Hawkey and Barker, 2004) and the more effective any assessment criteria and scales based on our descriptions will be. A sophisticated linguistic description of typical performance at each level would be able to define the linguistic characteristics that mark one level of performance from another. Such a description would also allow test developers to make descriptors more detailed. This would be well received by IELTS raters (Shaw, 2004, pp 6).

1.3 Research objectives

This study aims to document the linguistic markers of the different levels of English language writing proficiency defined by the academic version of the IELTS Writing module. The IELTS test (Academic Version) is administered in approximately 122 countries worldwide (http://www.ielts.org) and is used to assess the English language proficiency of non-native speakers of English who are planning to study at English-medium, tertiary-level institutions. The Academic Writing module, which is the focus of this study, is one of four modules (Listening, Reading, Writing and Speaking) and comprises two tasks (IELTS Handbook, 2005, pp 8-9). Test-takers are graded separately on both tasks using an analytic scale. Their final band for the Writing module is a weighted average of these two marks (where the second task is weighted more than the first).

Our original plan was to examine performances across all bands but performances at levels 1, 2 and 9 were not available so we examined scripts at band levels 3–8 only.

The central questions that the study addresses are:

1. What are the defining characteristics of written language performance at each IELTS band with regards to:
   a. frequency, type and function of cohesive devices used
   b. vocabulary richness
   c. syntactic complexity
   d. grammatical accuracy
2. How do these features of written language change from one IELTS level to the next across the 3–8 band range?
3. What are the effects of L1 and writing task type on the measures of proficiency under (1)?

We narrowed down and organised the target linguistic features in such a way as to cover a range of key areas of language and also to allow other users of this research to establish links with other frameworks, such as Cambridge ESOL's Common Scale for Writing (see Hawkey and Barker, 2004) and the Common European Framework of Reference for Languages (Council of Europe, 2001). We also take into consideration how the learners' first language and the type of task may affect their performances at different levels.

This report describes the completed study and discusses its main findings. It begins with an overview of the literature pertaining to analytic measures of L2 proficiency, previous research into the linguistic features that characterise different IELTS band levels and a discussion of potential intervening factors such as L1 and task effects. It then presents the design issues arising during the study and gives a full description of the final sample and the background data collected for each test-taker. Subsequent sections present the analyses for each target area of language: cohesive devices; vocabulary richness; syntactic complexity; and grammatical accuracy. The final section summarises and discusses the findings and their implications for further research.

2 LITERATURE REVIEW

The question of what characterises the written language of different IELTS band levels could be investigated in at least two ways. One approach would be to study writing descriptors and rater behaviour and perceptions (see McNamara, 1996, chapter 5 for examples of how this can be done). A second approach consists of investigating written performances that have been placed at different band levels with the aim of discovering the linguistic features that scripts placed at each level have in common. The present study adopts the second type of approach, building on previous work by Kennedy and Thorp (2002) and Mayor et al (2002).

2.1 Analytic measures of developing L2 proficiency

Larsen-Freeman (1978, pp 440) suggests that the ideal measure of linguistic ability should 'increase uniformly and linearly as learners proceed towards full acquisition of a target language'.
This preference seems justified, as proficiency scales are typically linear. However, the expectation that the rate of progress will be uniform within and across individuals and that all areas of language will make uniform progress is not justified by the research evidence available at present. For instance, the rate of L2 development can vary markedly from one individual to another (eg Perdue and Klein, 1993; Skehan, 1989; Slavoff and Johnson, 1995), and the close link between the development of a given property X and the subsequent development of another property Y that is typical of first language acquisition is not always found in second/foreign language acquisition (eg Clahsen and Muysken, 1986, 1989; Meisel, 1997). Therefore, we do not think that the requirement to increase uniformly is necessary, desirable or indeed defensible. Consequently, a more realistic pursuit would be

to look for the ideal group of measures that, when applied together, produced a learner language profile that could be reliably classified as being at a given level in a predetermined scale.

Wolfe-Quintero et al's thorough meta-study of fluency, accuracy and complexity measures of L2 writing proficiency (1998, pp 119) suggests a number of measures that could be profitably investigated:

- words per t-unit (see section 5.1 for a definition)
- words per clause
- words per error-free t-unit
- clauses per t-unit
- dependent clauses per clause
- word type measure
- sophisticated word type measure
- error-free t-unit per t-unit
- errors per t-unit.

2.2 Linguistic features characteristic of each IELTS band level

In this section we consider studies that have investigated measures of proficiency in a context more closely related to ours. These studies augment the selection of measures suggested by Wolf-Quintero et al (1998).

Mayor et al (2002, pp 46) found that the strongest predictors of band score in Writing Task 2 performances were the ones listed below.

- word count
- error rate
- complexity
- pattern of use of the impersonal pronoun 'one'.

Kennedy and Thorp (2002) confirmed these findings and found the following further trend:

- overt cohesive devices were used more frequently at IELTS levels 4 and 6 and less at levels 8 and 9, where cohesion was expressed more frequently through other means more generally in line with the native speaker norm; these findings are similar to those of Flowerdew (1998, cited in Kennedy and Thorp, 2002, pp 102).

The following were not good predictors of band score:

- type of theme (Mayor et al, 2002, pp 21)
- punctuation errors (Mayor et al, 2002, pp 6)
- number of t-units containing at least one dependent clause (Mayor et al, 2002, pp 14); this is at odds with the findings of Wolfe-Quintero et al's (1998) meta-study and deserves further investigation.

Despite the fact that these studies were able to identify some strong predictors of band level in their written performances, there seemed to be a complex network of interactions between some of the variables under investigation, and so the interpretation of their findings should not be oversimplified and generalised indiscriminately. In the next section (see 2.3, below) we discuss some of the potentially interacting variables that should not be ignored.

Hawkey and Barker (2004) also carried out a careful analysis of written performances at different levels with the aim of identifying features characteristic of each level. This study did not use IELTS band levels but rather the current FCE marking scheme and the 5 levels of the Cambridge ESOL Common Scale for Writing (CSW). This was applied to 108 FCE scripts, 113 CAE scripts and 67 CPE scripts (total of 288 scripts or 53,000 words). After a thorough rating procedure, the scripts that were unanimously placed at levels 2 (n = 8), 3 (n = 43), 4 (n = 18) and 5 (n = 29) of the scale were retained for further analysis (total of 98 scripts or 18,000 words).

Hawkey and Barker used the categories developed using an intuitive approach to the remarking of the 98 scripts in the subcorpus for proposing a new draft scale for writing. The criteria for identifying levels that they proposed after this intuitive marking process were based on the following groups of linguistic features:

- sophistication of language
- accuracy
- organisation and cohesion

The features that we have investigated relate directly to these categories, as shown in Table 2.1.

Hawkey and Barker (2004)/CSW features	**Features investigated in the present study**
Sophistication of language	Syntactic complexity Vocabulary richness
Accuracy	Grammatical accuracy
Organisation and cohesion	Cohesive devices

Table 2.1: Comparison of Hawkey and Barker (2004)/CSW target features and those in the present study

These features are present in the IELTS Academic Writing scales as Vocabulary and Sentence Structure (VSS) and Coherence and Cohesion/Communicative Quality (CC/CQ). In fact, these features seem to underpin several other proficiency and rating scales. For example, the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) scales are full of references to these key features. The reader can find evidence of how important these features are in this framework in the CEFR manual, for instance in the illustrative global scale (2001, pp 24) and the scales for overall written production (2001, pp 61), general linguistic range (2001, pp 110), vocabulary range (2001, pp 112), and grammatical accuracy (2001, pp 114).

2.3 Potential intervening factors

While the features mentioned above seem to be relatively good predictors of IELTS band score, it has been found that a number of other variables can affect the scores in different ways. This study addressed two of these potential intervening variables: L1 effect and task effect.

2.3.1 L1 effects

The role of the L1 on L2 development is well documented in the SLA literature (see Odlin 2003 for an overview), and the available evidence leads one to expect that the L1 will have some effect on specific L2 proficiency measures. It is therefore not surprising that L1 transfer has been found to have some clear and specific effects on L2 writing performance. For example, Mayor et al (2002) found that the L1 (Chinese vs Greek) affected Writing Task 2 performances in the following areas:

- complexity: this was measured as number of embedded clauses and the results showed that the L1 had significant effects on the type of clauses used by the learners, while band level did not make a significant difference (2002, pp 14)
- grammar errors: low-scoring Chinese L1 scripts had significantly more grammatical errors than comparable Greek L1 scripts (2002, pp 7 and 10)
- use of themes: L1 Chinese writers use more t-units and therefore more themes (2002, pp 25).

The writer's L1 did not seem to have an observable effect on the following:

- spelling errors (pp 7)
- punctuation errors (pp 7)
- preposition errors (pp 7)
- lexical errors (pp 7)
- overall number of errors (pp 7).

This study will make systematic analyses of possible L1 effects for each measure investigated.

2.3.2 Task effects

Mayor et al (2002) compared the performances of L1 Chinese and L1 Greek speakers on two versions of Writing Task 2 and found that the candidates' performances were similar on the two versions across levels overall. Nevertheless, some differences were found between the performances on each version of the test as follows:

- error frequency in different categories was comparable, except for preposition and lexis/idiom errors (2002, pp 10)
- number of t-units that included dependent clauses (2002, pp 14 and 47).

Unfortunately it was not possible to collect a balanced selection of test versions for the present study (primarily because we prioritised the variables band level and L1 over test version), so we will not conduct comparisons across different test versions. We will examine potential task effects by analysing Task 1 and Task 2 scripts separately and establish comparisons where relevant.

3 RESEARCH DESIGN

The purpose of the study was to explore the defining characteristics of written language performance at each IELTS band level with regard to cohesive devices used, vocabulary richness, syntactic complexity and grammatical accuracy. We were interested in how these features of written language change from one IELTS level to the next across the 3–8 band range and in the effects of L1 and writing task on the measures of proficiency we had selected.

Table 3.1 shows a general comparison of some key design features of the present study and some of the studies discussed in the previous section. The current study builds upon previous studies by looking at a much larger data set and at both the IELTS Academic Writing tasks. Like the Mayor et al (2002) study, it has controls for L1.

Study	No. of scripts	Corpus size (words)	IELTS band levels investigated	Writing Tasks	Versions of test	L1s
Present study	550†	**132,618**	**3 to 8**	1 and 2	26	Chinese and Spanish
Mayor et al (2002)	186	56,154	5 vs 7 and 8	2	2	Chinese and Greek
Kennedy and Thorp (2002)	130	35,464	4, 6, 8, 9 (8 and 9 conflated for analysis)	2	1	reported as unknown; presumably mixed
Hawkey and Barker (2004)	288	53,000	n/a; they were FCE, CAE and CPE	1	1	not reported; presumably mixed

†275 of these were Task 1 scripts and 275 were Task 2 scripts, a pair per learner

Table 3.1: Comparison of coverage of the present study and some previous studies

3.1 Sampling

We requested approximately equal numbers of scripts at each band level (1–9), balanced for L1 (50% L1 Chinese, 50% L1 Spanish). However, it was not possible to obtain scripts for band levels 1, 2 and 9 since these are much less common than the other levels in the current population of IELTS test takers. We received 159 scripts from centres across China and 116 scripts from four Latin American countries (Colombia, Mexico, Peru and Ecuador). Table 3.2 presents a summary of the different types of scripts that make up our corpus.

Band	L1 Chinese Centre	L1 Spanish Centre	Total
Band 9	0	0	0
Band 8	1	7	8
Band 7	15	33	48
Band 6	45	38	83
Band 5	53	29	82
Band 4	33	9	42
Band 3	12	0	12
Band 2	0	0	0
Band 1	0	0	0
Total scripts	159	116	275
Total no. of words	72,631	59,987	132,618

†The number of scripts in this table and in Figure 1 should be doubled if Task 1 and Task 2 are counted as separate scripts.

Table 3.2: Scripts in our corpus

Although the distribution of scripts by L1 and band is uneven, and therefore not ideal for some of the planned comparisons, the differences between the L1 Chinese centres and L1 Spanish centres regarding mark ranges and frequencies within each band are interesting in themselves. The data suggest that test-takers in centres in China tend to take the test when they are at lower levels of L2 proficiency than test-takers in centres in Latin America. We will not explore here the reasons behind these differences or the implications that such differences may have for Cambridge ESOL and other stakeholders, but it is nevertheless a fact worth mentioning.

3.2 Background data

In order to protect the anonymity of test-takers and to maintain high levels of test and test-performance security, test-takers' writing scripts and their responses to the Candidate Information Sheet (CIS) are stored separately. This data has to be reconciled by hand and, for this study, it has not been possible to complete the background information for every script in the data set. Table 3.3 presents the background data that we have been able to retrieve.

Background data		L1 Chinese Centre	L1 Spanish Centre	Total
Gender:	Male	59	53	112
	Female	55	51	106
First Language:	Chinese	128	-	128
	Spanish	-	113	113
Age:	16 – 25	87	50	137
	26 – 35	25	49	74
	36 or more	2	5	7
Years of L2 study:	less than 5	12	50	62
	6 or more	102	53	155

Table 3.3: Background data that is available for the data set

The background information indicates that the balance of male and female test-takers was almost equal, as was the balance between the two L1 groups (Chinese and Spanish). The sample was generally from young test-takers in the age group 16–25 with six or more years of L2 study.

3.3 Definition of performance level

The performance levels that have been adopted for this study are the band scores that were reported to the students on their official test report form. These scores have been subject to the standard quality control mechanisms in place for IELTS and described in some detail by Tony Green in a posting on LTEST-L, a discussion list for language testing professionals and researchers (LTEST-L, 24 January 2006). It is clear from this correspondence that all IELTS examiners undergo training and accreditation. Though double-marking is not performed on every script, a sample of scripts from every administration is double-marked to monitor rater standards.

As the discussion on LTEST-L has demonstrated, this practice is not held in high regard and it would be preferable if all scripts (particularly those being used for research and analysis) were double-rated. Therefore, it would have been desirable to re-rate the scripts in our sample to confirm the reliability of the scores if at all possible. However, this proved unfeasible on this occasion due to funding constraints and a lack of access to appropriately trained raters. Nevertheless, we would argue that it has been appropriate to carry out the analyses that follow (sections 4.0 – 7.0) using the 'live' scores for these represent the judgements of trained and monitored IELTS examiners and these are the judgements upon which decisions have been made about the candidates in our study.

3.4 Transcribing, coding and retrieval of information

The transcription of scripts was carried out by two transcribers. The transcribers first met with one of the investigators to familiarise themselves with the transcription conventions to be adopted. All three transcribed the same set of scripts, discussed any differences that were found in the transcriptions and then agreed on the details of the conventions. We also set up a system for recording transcription queries, in case we (or other researchers) wish to return to them at a later point.

We have used Wordsmith 3.0 for numerous standard quantitative analyses of the texts. We also explored two coding and retrieval tools. One was CLAN X (MacWhinney, 2000) and the other was Atlas.ti 5.0 (Muhr, 2004). We adopted the latter due to considerations of transcription and coding time required to use Atlas and the searches that the application allows us to do. Details of the analyses performed with Atlas will be presented in relevant sections below. As suggested by an independent reviewer, in future research we would like to use QDA Miner (Provalis) for this incorporates a concordancer tool.

4 COHESIVE DEVICES

The text features coherence and cohesion are important aspects of the IELTS rating scales. Raters are required to give an analytic score for *coherence and cohesion* as part of the process for marking Task 1 scripts and a judgement of text coherence is implicit in the Task 2 analytic scale (represented by *communicative quality*: impact on reader, effect on target reader) (Hawkey and Barker, 2004). It is therefore useful to analyse text features that contribute to cohesiveness and 'flow'. The taxonomy of cohesive ties developed by Halliday and Hasan (1976) has been particularly influential in this area so it was the starting point in our review of approaches to the analysis of coherence in written texts. We also discuss the analysis of anaphoric reference, specifically the frequency and patterns of use of the demonstratives: this, that, these and those.

4.1 Review of measures

Halliday and Hasan (1976) argue that cohesion within a text is established through five categories of cohesive ties: reference (also referred to as *anaphora*); ellipsis; substitution; conjunction; and lexis. A common approach in the analysis of cohesion is to analyse the frequency, form and context of use of connectors. These range from simple additive conjunctions such as *and* to contrastive devices such as *however* and *since*. We set this analysis aside from the outset on the grounds that recent research has revealed very little relationship between the overt use of linking words and test-taker performance (Ghazzoul, in progress). Kennedy and Thorp (2002) further suggest that test-takers at the lower IELTS band levels are more likely to use explicit linking devices than test-takers at higher IELTS band levels.

We also set aside the analysis of lexical cohesion owing to a lack of time to complete a full analysis of the scripts. We did, however, explore the analysis of ellipsis and substitution with a subset of 42 scripts from our corpus. This subset of texts will be referred to in connection with other analyses in this report and represents the full range of levels from the two language groups under investigation (ie IELTS scripts assessed at levels 4–8 from learners with Spanish as their first language and scripts rated at levels 3–8 from Chinese L1 candidates). The texts selected were also those that might be considered 'perfect' examples of their band levels since they received the same band score for each of the analytic categories as well as for the final band level. However, we found that it was difficult to be certain when ellipsis and substitution had been intentionally and correctly used, particularly at the lower IELTS band levels. Two examples are presented below:

> **L1 Chinese/ Band 4/ Task 1**
>
> The pie chart show that about world electricity production by energy source within Europe in 1997.
>
> As we can see. solid fuels are <u>most</u>. And then are Nuclear 20%, Gas 18%, Oil 10% Water 7%. at least is Other renewables.
>
> **L1 Spanish/ Band 4/ Task 1**
>
> The chart show that the principal exporter countries are EEUU and Canada with a value to $9.800 m approx. The <u>following</u> is Oceania with $3.800 m approx, but the different between North America (EEUU and Canada) and Oceania is big, this different is about $6.000 m approx.

The two underlined items might represent ellipsis and substitution. In the first extract, 'most' might be reconstructed as 'the most widely used energy sources'. In the second extract, 'following' might be reconstructed as 'the next largest exporter'. These items might have been intended to link the ideas together but they have not been used appropriately. This presents a number of problems, including:

1. How might an instance of ellipsis or substitution be identified and should the analysis take into account *possible* intention on the part of the writer?
2. Might an overly generous identification of ellipsis and substitution at the lower IELTS band levels inflate the measurement of these features at these levels and therefore skew the results?

The Halliday and Hasan (1976) framework, developed as it was from the analysis of native speaker texts, does not offer insights into how non-native speaker error might be accounted for in the analysis. We therefore abandoned this analysis in the expectation that we were likely to cover some key aspects of this lexicogrammatical ground using a less problematic methodology in our analyses of grammatical accuracy and syntactic complexity.

Analyses of anaphoric reference can include the use of personals (eg he, she, it, they, hers), demonstratives (eg this, these, that, those), and comparatives (eg same, similar, likewise, other). However, a preliminary analysis of this data set revealed that the category of demonstratives was the most promising. The analysis performed borrows its theoretical framework from Botley (2000) who investigated anaphora in written texts by first language speakers of English. Botley first identified each occurrence of the demonstratives – this, that, these and those – and then assigned them a 5-character code. Basing his work on Halliday and Hasan's (1976) categories, as well as deriving categories from his data (three corpora of English texts), Botley identified five distinctive features which he classified as:

- Recoverability of Antecedent – the degree of availability of each demonstrative's antecedent either directly (from the text), or indirectly (through the reader's understanding of the text)
- Direction of Reference – whether each demonstrative's antecedent appears in the text prior to the demonstrative (anaphorically) or following the demonstrative (cataphorically)
- Phoric Type – whether the relationship between each demonstrative and its antecedent was interpreted semantically (ie referential phoric type) or syntactically (ie substitutional phoric type)

- Syntactic Function – the function that each demonstrative fulfils in the sentence, either as the head of a noun phrase or as a noun modifier
- Antecedent Type – the form of each demonstrative's antecedent; whether it referred to a noun (phrase) or a clause.

These five main categories were further sub-divided to more precisely describe each occurrence, as in Table 4.1 below (reproduced from Botley, 2000).

Feature	**Value 1**	**Value 2**	**Value 3**	**Value 4**	**Value 5**
Recoverability of Antecedent	D (directly recoverable)	I (indirectly recoverable)	N (non-recoverable)	O (not-applicable, eg exophoric)	none
Direction of Reference	A (anaphoric)	C (cataphoric)	0 (not applicable, eg exophoric or deictic)	None	none
Phoric Type	R (referential)	S (substitutional)	0 (not applicable)	None	none
Syntactic function	M (noun modifier)	H (head noun)	0 (not applicable)	None	none
Antecedent Type	N (nominal antecedent)	P (propositional / factual antecedent)	C (clausal antecedent)	J (adjectival antecedent)	0 (no antecedent)

Table 4.1: Values assigned to each occurrence of the demonstratives in Botley's (2000) framework

It is important to note that there is currently no automatic or semi-automatic system available for the identification of anaphora (Botley and McEnery, 2000, pp 3). Therefore, all the analyses and annotation have to be performed manually. We applied Botley's framework in two stages. First the occurrences of the four demonstratives under discussion were manually annotated. This was performed on the whole data set. Then, the subset of 42 scripts described earlier was annotated more closely using the five distinctive features Botley identified (see Table 4.1, above) with the following provisos.

1. Coding the Recoverability of Antecedent depends to some degree on the reader's willingness or ability to make inferences and connections. An antecedent that to some readers may seem to be directly (or at least indirectly) recoverable, may not be recoverable at all to others, especially if they are unfamiliar with the genre or style of writing.

2. Although Botley often uses 'anaphora' as an umbrella term for all anaphoric phenomena, including cataphora (reference to an antecedent yet to be identified), in the coding for Direction of Reference the two terms are used in contradistinction to one another. In practice, however, no examples of cataphora were found in our sample.

3. Deciding whether the Phoric Type is referential or substitutional proved in these examples to be quite problematic. Additionally Botley (2000) found very few in his data. Thus it is does not seem unreasonable to conclude that this usage is quite rare, and perhaps not the most useful indicator of language competence. As well as being somewhat problematic, this category was not considered relevant to this study, and it was therefore not used.

4. It might be expected that determining the Antecedent Type should be relatively clear. However, when the actual data are examined, there are often overlaps found between propositional and clausal antecedents. It was therefore decided that only syntactic sub-categories should be recognised, leaving the sub-category 'propositional / factual antecedents' out.

4.2 Frequency of use of demonstratives (this, that, these, those)

The qualitative data analysis software, Atlas-ti was used to code the scripts for the occurrences of the demonstratives – this, that, these and those. The annotations were further refined by indicating whether the demonstrative had been used correctly (ie in its correct form – single or plural) or incorrectly. The final set of annotations was then extracted to an Excel file and transferred to SPSS for quantitative analysis. The following analyses were performed:

- the mean frequency of use of the demonstratives – this, that, these and those – by L1, task and IELTS band level (including standard deviations).
- the mean frequency of use of demonstratives as a whole by L1, task and IELTS band level.

We also used the right/wrong data to check whether the demonstratives had been used correctly on the whole and found that the occurrence of incorrect use of demonstratives was low for both groups: L1 Chinese test-takers (Task 1 = 8% and Task 2 = 5%); L1 Spanish test-takers (Task 1 = 11% and Task 2 = 10%).

The mean frequency of use of the demonstratives – this, that, these and those – (including standard deviations) according to L1 and IELTS band level for Tasks 1 and 2 are presented in Appendix 1. One interesting trend is the negligible use of the demonstrative 'those' by test-takers at all levels in both language groups. It also appears that L1 Spanish test-takers are more likely on the whole to use demonstratives than their L1 Chinese counterparts. The tables in Appendix 1 further indicate that the pattern of use of individual demonstratives is not clear and that there is a lot of variability in use within bands. We decided, therefore, to explore the patterns for overall use of demonstratives. They are presented in Table 4.2.

	L1 Chinese Means (SD)		L1 Spanish Means (SD)	
	Task 1	Task 2	Task 1	Task 2
Band 3 (N = 7/0)	1.00 (1.16)	1.43 (2.57)	-	-
Band 4 (N = 29/8)	1.21 (1.24)	3.41 (2.61)	4.63 (3.96)	2.87 (1.13)
Band 5 (N = 45/28)	1.67 (1.46)	2.76 (1.90)	3.14 (2.19)	3.64 (2.16)
Band 6 (N = 38/38)	2.21 (1.80)	3.82 (2.48)	2.89 (1.78)	4.53 (2.79)
Band 7 (N = 9/32)	2.78 (2.91)	3.22 (1.99)	2.84 (1.78)	3.94 (2.11)
Band 8 (N = 0/7)	-	-	4.00 (2.77)	4.43 (2.94)

Table 4.2: The mean frequency of use of demonstratives as a whole by L1, task and IELTS band level

Table 4.2 reveals some interesting trends in total demonstrative use:

- L1 Spanish speakers generally make more use of demonstratives in both tasks than L1 Chinese speakers (by approximately 50%). This is to be expected given the closer relationship of Spanish to English.
- The pattern of demonstrative use in Task 1 is different for the two L1 groups. Figure 4.1 (following) shows that demonstrative use by the L1 Chinese test-takers rises steadily in line with increases in the IELTS band level awarded while demonstrative use by the L1 Spanish test-takers declines in line with increases in the IELTS band level until IELTS band level 8 where there is a sharp increase again.
- Task 2 generates more demonstrative use than Task 1. This justifies the inclusion of two writing tasks to generate different uses of the language.
- L1 Chinese speakers are less likely to use demonstratives in Task 1 than Task 2.
- L1 Spanish writers display a different trend in demonstrative use in both tasks. In Task 1 L1 Spanish writers are less likely to use demonstratives as their IELTS band level increases whereas in Task 2 they are more likely to use demonstratives as their IELTS band level increases.

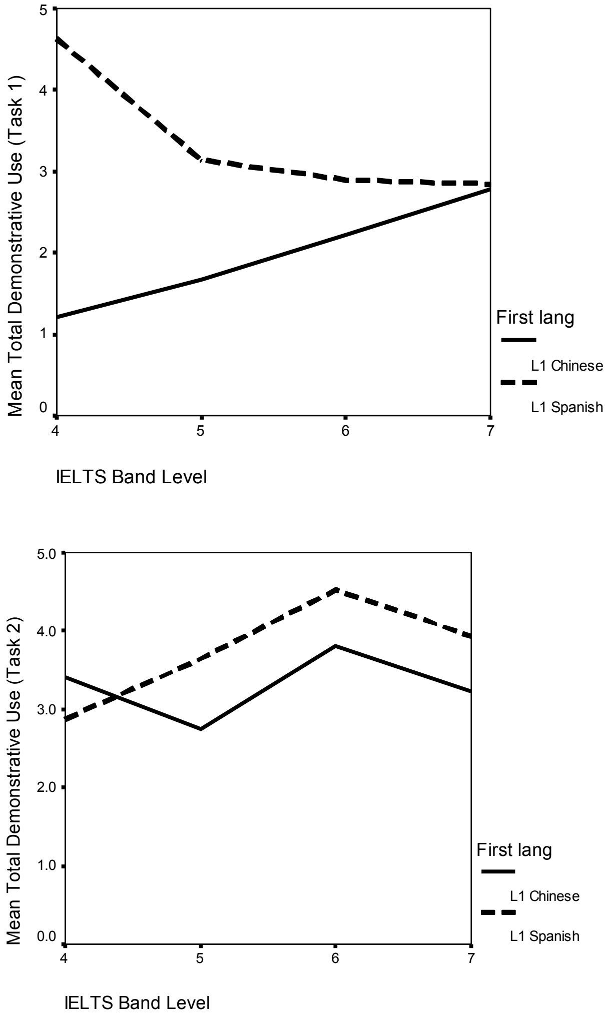


Figure 4.1: The mean frequency of use of demonstratives as a whole by L1, task and IELTS band level

We confirmed these patterns by doing a two-way ANOVA, between groups design (ie between L1 Spanish speakers and L1 Chinese speakers), to test the main effects and potential interactions of IELTS band level with L1 on the mean total demonstrative use for each task. The calculations covered only bands 4–7 because we lacked data from one L1 group for the other bands. Table 4.3 (following) shows the summary statistics for the two-way ANOVA band x L1, DV: mean total demonstrative use for Task 1.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	138.789(a)	7	19.827	5.569	.000
Intercept	1107.146	1	1107.146	310.960	.000
F_BAND	6.501	3	2.167	.609	.610
FIRST_LA	77.253	1	77.253	21.698	.000
F_BAND * FIRST_LA	46.238	3	15.413	4.329	.005
Error	779.731	219	3.560		
Total	2227.000	227			
Corrected Total	918.520	226			

a R Squared = .151 (Adjusted R Squared = .124)

Table 4.3: Summary statistics for the two-way ANOVA band x L1, DV: Mean total demonstrative use for Task 1

For Task 1 we were able to confirm that IELTS band level did not have a significant effect on the mean total demonstrative use. However, L1 did have a significant effect on the mean total demonstrative use (F = 21.698, p<0.01). There was also a significant interaction (F = 4.329, p<0.01) between IELTS band score and L1. This finding suggests that L1 Spanish speakers are more likely to use demonstratives at lower levels of language proficiency but that these are gradually replaced at higher levels of language proficiency by other types of cohesive tie (perhaps lexical ties). However, the steady increase in the use of demonstratives by L1 Chinese speakers suggests that this set of cohesive ties is not present at very low levels of language proficiency and is acquired as language proficiency increases.

However, it is important to note that the increase in demonstrative use by the L1 Chinese speakers does not overtake the demonstrative use by L1 Spanish speakers at the IELTS band levels studied. It would be very interesting to analyse a matched data set at IELTS band levels 8 and 9 in order to check whether the differences in demonstrative use level out at higher levels of language proficiency. Additionally, this pattern is not replicated in Task 2 where there is a generally higher use of demonstratives overall. Figure 4.1 shows a linear increase in demonstrative use until IELTS band level 6 for both L1 groups after which mean frequency of use of demonstratives begins to tail off, suggesting that other cohesive ties come into use at higher levels of language proficiency.

Table 4.4 shows the summary statistics for the two-way ANOVA band x L1, DV: mean total demonstrative use for Task 2. We were able to confirm that L1 did not have a significant effect on the mean total demonstrative use and also that there was no interaction between IELTS band score and L1. However, the effect of IELTS band level on the mean total demonstrative use approached significance (F = 2.562, p<0.056). Post-hoc tests (Tukey HSD) showed that the effect was due to a significant difference in demonstrative use between IELTS band levels 5 and 6. This finding cannot be explained with the data available. Among the possible explanations that will need to be considered is whether the boundary between IELTS level 5 and 6 marks a step in the language acquisition process and results in the temporary loss of some gains made at earlier stages.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	76.622(a)	7	10.946	2.050	.050
Intercept	1926.767	1	1926.767	360.878	.000
F_BAND	41.021	3	13.674	2.561	.056
FIRST_LA	7.634	1	7.634	1.430	.233
F_BAND * FIRST_LA	9.871	3	3.290	.616	.605
Error	1169.264	219	5.339		
Total	4208.000	227			
Corrected Total	1245.885	226			

a R Squared = .061 (Adjusted R Squared = .032)

Table 4.4: Summary statistics for the two-way ANOVA band x L1, DV: Mean total demonstrative use for Task 2

We also compared the distribution of each type of demonstrative in the two learner groups at each band level to a corpus of English native speaker performances on IELTS test tasks (see the native speaker section of LANCAWE, http://www.ling.lancs.ac.uk/groups/slarg/lancawe/). We found that while the L1 Spanish group gradually approached the distribution of English native speakers as band level increased, the opposite was true of the L1 Chinese group. A qualitative investigation of this phenomenon would help to identify possible causes for the contradictory developmental trend in the two learner groups, but we are able to conclude at this point that there is an effect of the L1 in the distribution of demonstratives used in both Writing tasks.

4.3 Use of demonstratives (this, that, these, those)

An investigation of the usage of demonstratives (this, that, these and those) was carried out using Botley's (2000) five features of demonstratives:

- Recoverability of Antecedent
- Direction of Reference
- Phoric Type
- Syntactic Function
- Antecedent Type.

Since the annotation of the data set had to be done manually and this was an extremely slow process, we restricted our initial analysis to the subset of 42 scripts described in 4.1 (above) in order to establish whether the results added to the findings from our analysis of the frequency of use of the demonstratives.

4.3.1 Recoverability

As expected, there was a high degree of recoverability of the antecedents of demonstratives. Of the 124 occurrences, 103 were directly recoverable by the reader/researcher. Fourteen were indirectly recoverable, and only seven were deemed non-recoverable, or exophoric. The classification of the 124 occurrences according to their recoverability is shown in Table 4.5. Bearing in mind that L1 Chinese speakers do not use demonstratives as frequently as L1 Spanish speakers, L1 does not appear to influence the pattern of recoverability of references.

	Directly recoverable			Indirectly recoverable			Non-recoverable		
L1	Chinese	Spanish	All	Chinese	Spanish	All	Chinese	Spanish	All
level 3	6	/	6	2	/	2	0	/	0
level 4	8	17	25	2	0	2	1	2	3
level 5	7	11	18	2	2	4	0	0	0
level 6	4	12	16	2	1	3	1	1	2
level 7	7	8	15	1	1	2	0	1	1
level 8	8	15	23	0	1	1	0	1	1
total	40	63	104	9	5	14	2	5	7

Table 4.5: Recoverability of antecedents by level and L1

4.3.2 Direction of reference

The majority of the occurrences of demonstratives in the samples were anaphoric, referring to antecedents earlier in the text. Some antecedents could not be identified, however, because of the exophoric or deictic nature of the reference. Table 4.6 shows the distribution of these categories by L1 and by level. As in the case of recoverability, L1 does not appear to influence the pattern of direction of reference.

	anaphoric reference			cataphoric reference			exophoric/deictic reference		
L1	Chinese	Spanish	All	Chinese	Spanish	All	Chinese	Spanish	All
level 3	8	/	8	0	/	0	0	/	0
level 4	10	17	27	0	0	0	1	2	3
level 5	8	13	21	0	0	0	1	0	1
level 6	6	13	19	0	0	0	1	1	2
level 7	8	9	17	0	0	0	0	1	1
level 8	8	15	23	0	0	0	0	2	2
total	48	67	115	0	0	0	3	6	9

Table 4.6: Direction of reference of antecedents by level and L1

4.3.3 Syntactic function

The number of occurrences of demonstratives used anaphorically as noun modifiers accounted for the vast majority of the total number, as shown in Table 4.7. L1 does not appear to influence the pattern in the syntactic function of the demonstratives.

	noun modifier			head noun		
L1	Chinese	Spanish	All	Chinese	Spanish	All
level 3	8	/	8	0	/	0
level 4	3	17	20	8	2	10
level 5	3	8	11	6	5	11
level 6	4	13	17	3	1	4
level 7	7	4	11	1	6	7
level 8	7	13	20	1	4	5
Total	32	55	87	19	18	37

Table 4.7: Syntactic function of demonstratives by level and L1

4.3.4 Antecedent type

The majority of antecedents referred to in these texts were nominals, either single nouns or noun phrases. It was also quite common for writers to refer to propositions earlier in the text, especially those that were responses to statements or arguments and were expressed as clauses. There were no adjectival antecedents, so they do not appear in Table 4.8. The distributions of antecedent type by L1 and level are shown below.

	Nominal			Clausal			No Antecedent		
L1	ch	Sp	All	ch	sp	All	Ch	sp	All
level 3	7	/	7	1	/	1	0	/	0
level 4	6	14	20	4	3	7	1	2	3
level 5	1	6	7	7	6	13	1	1	2
level 6	4	8	12	2	5	7	1	1	2
level 7	6	3	9	2	6	8	0	1	1
level 8	5	7	12	3	8	11	0	2	2
total	29	38	67	19	28	47	3	7	10

Table 4.8: Antecedent type by level and L1

The analysis of this carefully selected subset suggested that a wider analysis would be unlikely to be revealing.

4.4 Summary of findings

Our analysis of cohesive devices was based on Halliday and Hasan's (1976) framework of cohesive ties. We have explored the use of anaphoric reference in the form of the demonstratives (this, that, these and those). Our analysis revealed the following.

1. L1 Spanish speakers use approximately 50% more demonstratives than L1 Chinese speakers.
2. Test-takers are more likely to use demonstratives when responding to IELTS Task 2 than to IELTS Task 1.
3. Demonstrative use by both L1 groups appears to be influenced by the task (ie the type of writing) that they are asked to do. However, the influence is different in nature. For L1 Chinese speakers the task affects the number of demonstratives used but the relationship between demonstrative use and IELTS band level remains the same. For L1 Spanish speakers, the number of demonstratives used is fairly stable but the relationship between demonstrative use and IELTS band level differs from Task 1 to Task 2.
4. Use of demonstratives appears to tail off at higher levels of language proficiency, suggesting that other cohesive ties come into use. We would suggest that writers at higher IELTS band levels are more likely to use lexical ties to create cohesion. We would therefore expect performances at higher IELTS band levels to display greater lexical variation and sophistication.

It is important to note that the analysis of cohesive devices highlights the difficulties of applying measures developed for the analysis of L1 texts to L2 performances. Adapting these measures to the study of L2 performances is an important area for future research.

5 VOCABULARY RICHNESS

The text features associated with lexical richness figure prominently in the IELTS rating scales. Raters are required to give an analytic score for *vocabulary and sentence structure* as part of the process for marking both Task 1 and Task 2. Since the size of our sample at each IELTS band level varies considerably, we have calculated the mean scores at each band level for measures of lexical output, lexical variation, lexical density and lexical sophistication in order to account for differences in N size between groups.

5.1 Review of measures

The simplest measure of vocabulary richness available looks at lexical output, counting first the total number of words (tokens) written and then the total number of *different word forms* (types) written. Measures of lexical output calculate the number of tokens and types and the results can be correlated against language proficiency as reflected in IELTS band scores. The results of this analysis are presented in 5.2.

The analysis of lexical output can be taken a step further by examining the ratio of types to tokens. This is a measure of lexical variation/diversity because the higher the number of types in relation to tokens the more varied/diverse the vocabulary used. The traditional approach to calculating lexical variation has been the Type-Token Ratio (TTR) where the number of types is divided by the number of tokens and multiplied by 100. However, this approach is affected by text length so a more robust method of measuring lexical variation has been developed by Malvern and Richards (2002) (see also Duran et al, 2004), a D-value. Meara and Miralpeix (2004) have developed software to calculate the D-value for texts. However, we did not learn of the availability of these tools until late in the project and the calculations were not possible in the time available. Instead, in 5.3 we present the results of our analysis of lexical variation using the more traditional TTR approach.

A third and even closer examination of the nature of vocabulary use in a text is the measure of lexical density (Ure, 1971). This measure compares the number of lexical words used to the number of grammatical words. We have adopted O'Loughlin's (2001) definition of grammatical and lexical items as well as his division of lexical items into high-frequency and low-frequency items and, as suggested by Halliday (1985, pp 64-5, cited in O'Loughlin, 2001, pp 102), weighted the high-frequency items at half the value of the low-frequency items because this is likely to provide a truer estimate of lexical density. To identify high-frequency items in our data set we used the lexical items found in the 750 most frequently occurring words in the British National Corpus (BNC). In 5.4 we present the results of our analysis of 'weighted' lexical density.

Measures of lexical sophistication provide insight into the number of unusual or rare words used by a writer. We adopted the approach developed by Nation and Heatley (1996), using the Range program. The program classifies the words in a text into four categories. The first two categories are the first and second thousand most frequently occurring words in English (West, 1953). The third category is the Academic Word List (Coxhead, 2000) and contains 570 word families. The final category is an open category for all words that are not contained in the first three lists. We present our analyses of lexical sophistication in 5.5.

We also considered but did not implement an approach developed by Engber (1995) which calculates the percentage of lexical errors in a text. Engber developed a classification of lexical errors as one measure in an investigation of the relationship between lexical richness and the quality of ESL compositions. She found that her measure of *error-free lexical variation* correlated best with the students' scores. Any application of this approach would have to take into consideration Laufer and Nation's (1995) criticism that it does not distinguish between errors in types and tokens. There might therefore be considerable 'double-counting' in the calculation of lexical error. Other criticisms of the approach are that it is sometimes difficult to distinguish between lexical and grammatical errors and that Engber's framework does not take into account the relative seriousness of different errors (Read, 2000, pp 205).

5.2 Lexical output

An analysis of lexical output looks simply at the number of words produced by the test-takers. We performed the following calculations:

- mean number of tokens and types for the whole data set (including standard deviations)
  by task and IELTS band level
- mean number of tokens and types for each L1 group (including standard deviations) by IELTS band level.

Table 5.1 presents the mean number of tokens and types (including standard deviations) for the whole data set, according to IELTS band level for Task 1 and Task 2. The results show that, as might be expected, the mean number of tokens and types increases with the test-takers' IELTS band level. This indicates that more proficient test-takers are likely to produce more words (and more *different* words – types) than less proficient test-takers.

	Task 1 Means (SD)		Task 2 Means (SD)	
	Tokens	Types	Tokens	Types
Band 3 (N = 12)	138.2 (38.9)	53.1 (13.3)	132.3 (52.2)	65.7 (19.4)
Band 4 (N = 42)	163.6 (54.8)	66.2 (20.8)	253.0 (70.6)	127.0 (32.6)
Band 5 (N = 82)	189.1 (50.7)	76.3 (24.7)	284.0 (56.8)	137.0 (25.3)
Band 6 (N = 83)	208.6 (46.1)	86.4 (22.2)	308.6 (54.6)	152.3 (25.6)
Band 7 (N = 48)	223.6 (51.8)	94.2 (20.6)	312.2 (56.8)	159.0 (25.6)
Band 8 (N = 8)	230.6 (39.1)	102.6 (26.5)	323.3 (31.5)	160.4 (12.9)

Table 5.1: Lexical output for the whole data set, according to IELTS band level for Task 1 and 2

However, the large Standard Deviation figures for both tokens and types indicate a wide variation in lexical output within each band level. To check whether this is accounted for by the test-takers' L1, we split the data set into the two L1 groups (L1 Chinese and L1 Spanish) and recalculated the figures. Tables 5.2 and 5.3 present the total number of tokens and types and the mean number of tokens and types (including standard deviations) for each L1 (Chinese and Spanish), according to IELTS band level for Task 1 and Task 2.

	L1 Chinese Means (SD)		L1 Spanish Means (SD)	
	Tokens	Types	Tokens	Types
Band 3 (N = 7/0)	142.4 (42.1)	53.6 (15.2)	-	-
Band 4 (N = 29/8)	150.2 (43.6)	60.2 (16.8)	203.8 (77.7)	86.1 (26.0)
Band 5 (N = 45/28)	173.8 (38.2)	66.1 (13.6)	218.1 (61.6)	94.3 (30.6)
Band 6 (N = 38/38)	190.8 (34.1)	76.0 (12.5)	224.0 (52.5)	96.2 (26.1)
Band 7 (N = 9/32)	190.2 (47.8)	83.0 (15.2)	240.7 (49.7)	99.4 (22.1)
Band 8 (N = 0/7)	-	-	237.3 (37.0)	103.4 (28.6)

Table 5.2: Lexical output for L1 Chinese and L1 Spanish scripts, according to IELTS band level for Task 1

	L1 Chinese Means (SD)		L1 Spanish Means (SD)	
	Tokens	Types	Tokens	Types
Band 3 (N = 7/0)	128.3 (41.3)	65.3 (14.4)	-	-
Band 4 (N = 29/8)	259.4 (70.7)	132.0 (32.4)	211.4 (53.5)	105.1 (21.5)
Band 5 (N = 45/28)	290.4 (49.2)	139.5 (24.3)	276.8 (73.4)	135.6 (26.5)
Band 6 (N = 38/38)	308.6 (43.6)	156.2 (20.4)	306.4 (66.4)	146.8 (29.4)
Band 7 (N = 9/32)	323.6 (51.3)	174.9 (12.2)	310.3 (63.3)	153.5 (28.5)
Band 8 (N = 0/7)	-	-	321.7 (33.7)	159.3 (13.5)

Table 5.3: Lexical output for L1 Chinese and L1 Spanish scripts, according to IELTS band level for Task 2

Both tables show that the pattern of high standard deviations persists in both L1 groups. Additionally, in Task 1 the lexical output for the L1 Chinese group drops off slightly between IELTS band 6 and band 7 and for the L1 Spanish group, it drops off between IELTS band 7 and 8. It is possible that the slight dip in the figures can be explained by the drop in sample size at the top end of the scale for both L1 groups.

This small anomaly aside, there appears to be a link between band level and the mean number of types and tokens produced. We tested this by calculating Pearson correlations between band score and mean number of types and tokens. We found that the Pearson correlation between band score and mean number of types per script was moderate for Task 1 and Task 2 scripts, although slightly higher for the latter; both were significant (Task 1: r=0.442, n=278, p<0.01; Task 2: r=0.529, n=278, p<0.01). We also found that the Pearson correlation between band score and mean number of tokens per script was moderate for Task 1 and Task 2 scripts, both were significant, and slightly weaker than for mean number of types (Task 1: r=0.420, n=278, p<0.01; Task 2: r=0.472, n=278, p<0.01). These findings are in line with evidence in the literature, as other studies have also found moderate correlations between number of tokens produced and proficiency level (eg Homburg, 1984; Larsen-Freeman, 1978; Teddick, 1990).

We also performed a two-way ANOVA to test the main effects and potential interactions of the factors band level x L1, with the mean number of types per script as dependent variable. We excluded band levels 3 and 8 as we lacked data from one L1 group in each case so the analysis only covers band levels 4–7. We found that there were significant effects of both band level and L1 on mean type scores both for Task 1 and Task 2, but the interaction between these two factors was not significant in either case. Post-hoc tests (Tukey HSD) showed that the different band levels were significantly different except for levels 6 and 7 in Tasks 1 and 2 and levels 4 and 5 in Task 2. The following tables summarise the results.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	46497.544(a)	7	6642.506	16.013	.000
Intercept	1310182.636	1	1310182.636	3158.445	.000
BANDF	7469.146	3	2489.715	6.002	.001
L1	20980.706	1	20980.706	50.578	.000
BANDF * L1	615.758	3	205.253	.495	.686
Error	103704.708	250	414.819		
Total	1848615.000	258			
Corrected Total	150202.252	257			

Table 5.4: Summary statistics for two-way ANOVA band x L1, DV: mean types (Task 1)

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	44936.287(a)	7	6419.470	9.465	.000
Intercept	3906266.225	1	3906266.225	5759.686	.000
BANDF	41687.828	3	13895.943	20.489	.000
L1	10754.891	1	10754.891	15.858	.000
BANDF * L1	2773.152	3	924.384	1.363	.255
Error	169552.058	250	678.208		
Total	5617789.000	258			
Corrected Total	214488.345	257			

Table 5.5: Summary statistics for two-way ANOVA band x L1, DV: mean types (Task 2)

We repeated these procedures to test the potential main effects and interactions of band level x L1, with the mean number of *tokens* per script as dependent variable and found very similar results to the type analysis. There were significant effects of both band level and L1 on mean token scores both for Task 1 and Task 2, but the interaction between these two factors was not significant in either case. Post-hoc tests (Tukey HSD) showed that the different band levels were significantly different except for levels 6 and 7 in Tasks 1 and 2. A brief summary of the results is presented in Tables 5.6 and 5.7 on the following page.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	190632.209(a)	7	27233.173	12.593	.000
Intercept	7646635.208	1	7646635.208	3536.043	.000
BANDF	35647.954	3	11882.651	5.495	.001
L1	81206.449	1	81206.449	37.552	.000
BANDF * L1	4048.196	3	1349.399	.624	.600
Error	540620.896	250	2162.484		
Total	10830447.000	258			
Corrected Total	731253.105	257			

Table 5.6: Summary statistics for two-way ANOVA band x L1, DV: mean tokens (Task 1)

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	133598.670(a)	7	19085.524	5.554	.000
Intercept	15861476.323	1	15861476.323	4615.457	.000
BANDF	128755.243	3	42918.414	12.489	.000
L1	15015.888	1	15015.888	4.369	.038
BANDF * L1	7804.048	3	2601.349	.757	.519
Error	859149.908	250	3436.600		
Total	23106059.000	258			
Corrected Total	992748.578	257			

Table 5.7: Summary statistics for two-way ANOVA band x L1, DV: mean tokens (Task 2)

These results suggest that the number of words produced and the number of *different* words (types) produced does contribute to a test-taker's IELTS band level.

5.3 Lexical variation

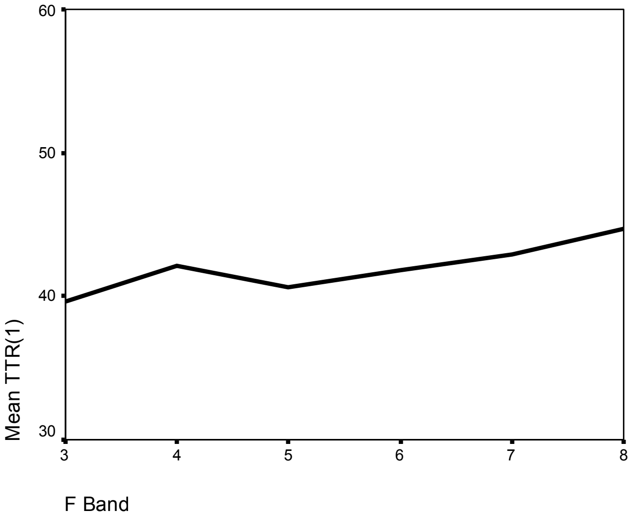
An analysis of lexical variation explores the number of different words (types) produced by the test-takers in relation to the total number of words produced. We performed the following calculations:

- mean Type-Token ratio (TTR) for the whole data set (including standard deviations) by task and IELTS band level
- mean TTR for each L1 group (including standard deviations) by IELTS band level.

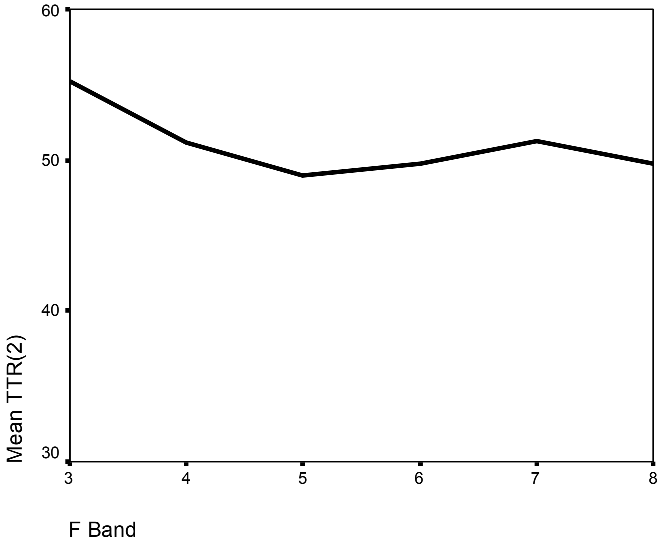
Table 5.8 presents the mean TTR (including standard deviations) for the whole data set, according to IELTS band level for Task 1 and Task 2 as well as the maximum and minimum TTR at each level. The results reveal a rather mixed picture. For Task 1 the pattern is generally linear (see Figure 5.1) with the mean TTR rising in line with the test-takers' IELTS band level. However, the mean TTR is higher at IELTS band level 4 than at IELTS band levels 5 or 6, producing a spike in the line.
This might be explained (at least partially) by the standard deviation and the range for this band level. The standard deviation figures are generally high but the performances at IELTS band level 4 have the highest standard deviation (10.1) and the range is 52, indicating much more variability in performance at this band level than at the others.

	Task 1 TTR			**Task 2 TTR**		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 12)	39.6 (9.2)	56.6	31.2	55.2 (17.1)	90.5	28.3
Band 4 (N = 42)	42.1 (10.1)	80.0	28.0	51.1 (7.0)	71.1	36.6
Band 5 (N = 82)	40.6 (7.8)	59.3	20.5	48.9 (6.9)	66.1	27.8
Band 6 (N = 83)	41.8 (7.3)	62.3	22.9	49.7 (5.4)	60.2	31.8
Band 7 (N = 48)	42.9 (6.6)	56.3	27.6	51.3 (5.1)	64.2	42.7
Band 8 (N = 8)	44.7 (9.0)	52.7	30.6	49.8 (3.5)	55.0	42.7

Table 5.8: Mean Type-Token ratio (TTR) for the whole data set (including standard deviations) by task and IELTS band level



Mean TTR: Task 1



Mean TTR: Task 2

Figure 5.1: Mean Type-Token ratio (TTR) for the whole data set by IELTS band level

Both graphs are broadly linear and kinks in the curve might be explained (at least partially) by the high standard deviation figures which indicate a high degree of variability within the bands. Also, since the TTR calculation is affected by the length of the texts, it is likely that it is also affected by the number of samples at each band level, particularly at IELTS band levels 3 and 8 where our sample sizes were particularly small.

The difference in the pattern between Task 1 and Task 2 indicates also that there is a slight task effect on the TTR calculation. It is also possible that the calculations have been affected by L1.

Tables 5.9 and 5.10 present the mean TTR (including standard deviations) for each L1 (Chinese and Spanish), according to IELTS band level for Task 1 and Task 2 as well as the maximum and minimum TTR at each level. They indicate that there is no clear relationship between band level and TTR for either task in either L1 group. We confirmed this by doing a two-way ANOVA, between groups design, to test the main effects and potential interactions of IELTS band level with L1 on the mean TTR for each task. Once again we excluded band levels 3 and 8 because we lacked data from one L1 group in each case.

	L1 Chinese TTR			L1 Spanish TTR		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 7/0)	38.3 (8.6)	56.6	31.2	-	-	-
Band 4 (N = 29/8)	41.8 (10.3)	80.0	28.6	44.3 (10.8)	65.1	31.8
Band 5 (N = 45/28)	38.8 (7.5)	54.7	23.6	43.2 (7.3)	59.3	31.0
Band 6 (N = 38/38)	40.3 (5.5)	55.6	28.6	43.6 (8.6)	62.3	22.9
Band 7 (N = 9/32)	44.9 (7.5)	52.8	29.1	41.8 (6.9)	56.3	27.6
Band 8 (N = 0/7)	-	-	-	43.5 (9.1)	51.4	30.6

Table 5.9: Mean Type-Token ratio (TTR) for L1 Chinese and L1 Spanish scripts according to IELTS band level for Task 1

	L1 Chinese TTR			L1 Spanish TTR		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 7/0)	54.2 (15.0)	72.3	28.3	-	-	-
Band 4 (N = 29/8)	51.9 (6.6)	71.1	41.2	50.9 (9.0)	64.0	36.6
Band 5 (N = 45/28)	48.4 (6.0)	60.8	36.9	50.3 (7.3)	66.1	36.7
Band 6 (N = 38/38)	50.8 (4.3)	58.2	42.9	48.5 (6.4)	60.2	31.8
Band 7 (N = 9/32)	54.7 (5.6)	60.6	42.7	49.8 (3.8)	59.7	42.7
Band 8 (N = 0/7)	-	-	-	49.7 (3.7)	55.0	42.7

Table 5.10: Mean Type-Token ratio (TTR) for L1 Chinese and L1 Spanish scripts according to IELTS band level for Task 2

Table 5.11 shows the summary statistics for the two-way ANOVA band x L1, DV: TTR for Task 1 and Table 5.12 shows the summary statistics for the two-way ANOVA band x L1, DV: TTR for Task 2.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	797.081(a)	7	113.869	1.858	.078
Intercept	278212.899	1	278212.899	4539.252	.000
F_BAND	142.479	3	47.493	.775	.509
FIRST_LA	121.835	1	121.835	1.988	.160
F_BAND * FIRST_LA	295.479	3	98.493	1.607	.189
Error	13422.614	219	61.290		
Total	407941.729	227			
Corrected Total	14219.695	226			

a R Squared = .056 (Adjusted R Squared = .026)

Table 5.11: Summary statistics for the two-way ANOVA band x L1, DV: TTR for Task 1

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	540.591(a)	7	77.227	2.210	.035
Intercept	398470.106	1	398470.106	11404.200	.000
F_BAND	221.035	3	73.678	2.109	.100
FIRST_LA	98.121	1	98.121	2.808	.095
F_BAND * FIRST_LA	291.873	3	97.291	2.784	.042
Error	7652.001	219	34.941		
Total	576696.036	227			
Corrected Total	8192.593	226			

a R Squared = .066 (Adjusted R Squared = .036)

Table 5.12: Summary statistics for the two-way ANOVA band x L1, DV: TTR for Task 2

We found that IELTS band level and L1 did not have a significant effect on the TTR for either Task 1 or Task 2. Additionally, for Task 1, there was no interaction between the two factors. However, there was a significant interaction ($F = 2.784$, $p<0.05$) between IELTS band score and L1 for Task 2. This indicates that the effect of the L1 (L1 Chinese or L1 Spanish) on the TTR was different for different IELTS band scores. Figure 5.2 shows the boxplot for the two-way ANOVA band x L1, DV: TTR (Task 2). It appears that the effect of the L1 is greater at IELTS band scores 6 and 7. This is in line with our expectation that L1 will have some effect on certain L2 proficiency measures. The effect is discernible for Task 2, the more open of the two tasks, and suggests that L1 exerts a stronger effect when the task is less controlled (and perhaps less supported) by the input material.

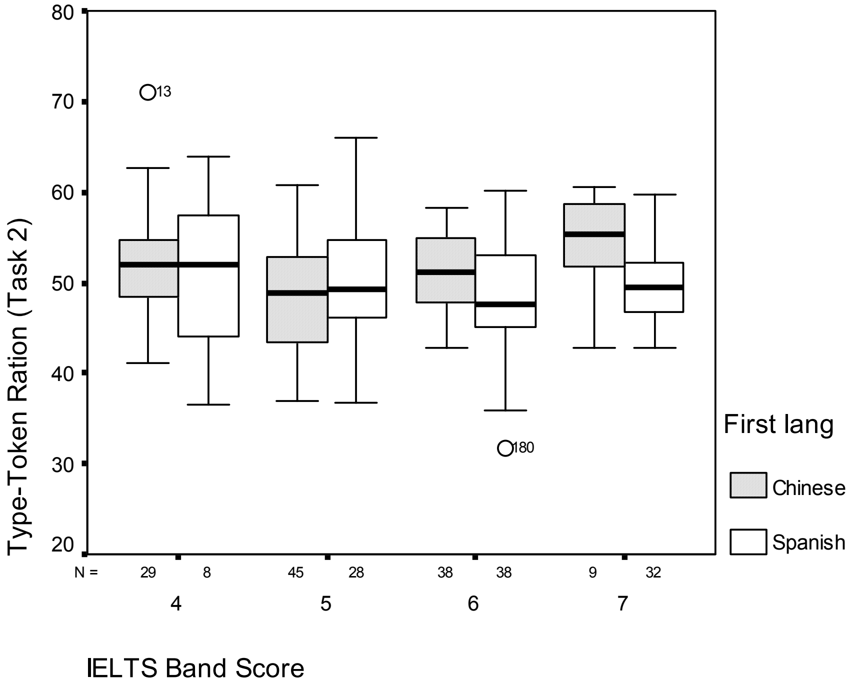


Figure 5.2: Boxplot for two-way ANOVA band x L1, DV: TTR (Task 2)

5.4 Lexical density

Measures of lexical density calculate the proportion of lexical words to grammatical words in the text and can also explore the extent to which test-takers at different IELTS band levels use low-frequency lexical words.

We adopted O'Loughlin's (2001) method for calculating lexical density. Items were allocated codes based on their function – grammatical or lexical. The COBUILD English Grammar reference book was consulted in the process of building lists for coding (eg lists of prepositions).

Words that were coded as grammatical items included:

- all forms of 'to be', 'to have' and 'to do', plus modals
- determiners, including quantifiers
- pro-forms
- interrogative adverbs
- contractions
- prepositions
- discourse markers.

All other words were coded as lexical items. We subdivided the lexical items into two groups – common lexical items and items that are not frequently used in English, using the BNC to identify high-frequency words.

We were aware that corpus frequencies, taken alone, give a very limited picture of a word's distribution in a corpus. As well as varying in raw frequency, words vary in the extent to which they are equally spread across the documents on the corpus. This 'burstiness' can be measured in a variety of ways (Church and Gale, 1995). One straightforward possibility is to take a large number of documents, all of the same length; count the frequency of a word in each of these documents, and calculate the (mean and) variance of this frequency and this is the approach adopted in this study.

The first 5,000 words of all documents (= files) longer than 5,000 words in the written part of the BNC were taken. A frequency list was produced for each of these (truncated) documents. From this we identified the 750 most frequently occurring words and categorised the lexical items in our data set as follows:

- Hlex: high frequency lexical items; those (non-grammatical, single word) items which appear in the top 750 of the BNC corpus frequency lists, plus their multi-word derivations (eg see, to see, have seen etc.)
- Llex: low frequency lexical items which did not appear in the BNC top 750. In contrast to O'Loughlin (2001) we allocated verbs in the infinitive form (with 'to') to whichever category they would have been in without 'to' since this made most intuitive sense.

In an attempt to account for the effect of topic upon the words produced, we created an additional category of Hrep. This was a subset of Llex and included items (plus their multi-word derivations) which did not appear in the BNC top 750, but which were frequent in this sample. In order to establish a cut-off point for this group of words, we adopted the frequency figure for the 750th word in the BNC top 750.

Decisions made during coding included the following.

1. Where items had two uses, but one was grammatical and the other lexical, the item was first coded for both. Then each occurrence was checked manually to correct the coding in context. For example the item 'past' was coded as either a grammatical item (preposition) or a lexical item (Llex) according to context.
2. The verbs 'to be', 'to have' and 'to do' were allocated a grammatical coding in all their forms, including the infinitive forms.
3. Contractions (eg 'can't') were coded as one item but counted as two in the calculations. This had to be done manually to ensure that they were recognised by the software.
4. Multiple word items were coded as one item (eg 'United Kingdom' which was coded separately from 'united' or 'kingdom'). The coding was checked to ensure that all items were coded, but that no item was coded twice.

The 50 most frequently occurring words in the L1 Chinese and L1 Spanish texts are presented in Appendix 2. Once the coding was complete, we performed the following calculations:

- mean lexical density for the whole data set (including standard deviations) by task and IELTS band
- mean lexical density for each L1 group (including standard deviations) by task and IELTS band.

We weighted the calculations by counting each of the high frequency lexical items as equal to half of the value of the low frequency lexical items (Halliday, 1985).

Table 5.13 presents the mean lexical density for the whole data set (including standard deviations) by task and IELTS band and Figure 5.3 presents the trend (or lack thereof) as a line graph.

	Task 1			Task 2		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 12)	1.00 (0.43)	1.95	0.50	0.70 (0.17)	1.09	0.46
Band 4 (N = 42)	0.76 (0.18)	1.13	0.46	0.67 (0.12)	0.99	0.41
Band 5 (N = 82)	0.85 (0.24)	1.59	0.45	0.66 (0.10)	0.92	0.42
Band 6 (N = 83)	0.79 (0.18)	1.48	0.49	0.69 (0.12)	1.08	0.37
Band 7 (N = 48)	0.82 (0.15)	1.23	0.55	0.75 (0.15)	1.15	0.49
Band 8 (N = 8)	0.82 (0.21)	1.95	0.45	0.69 (0.13)	1.15	0.37

Table 5.13: Mean lexical density for the whole data set (including standard deviations) by task and IELTS band

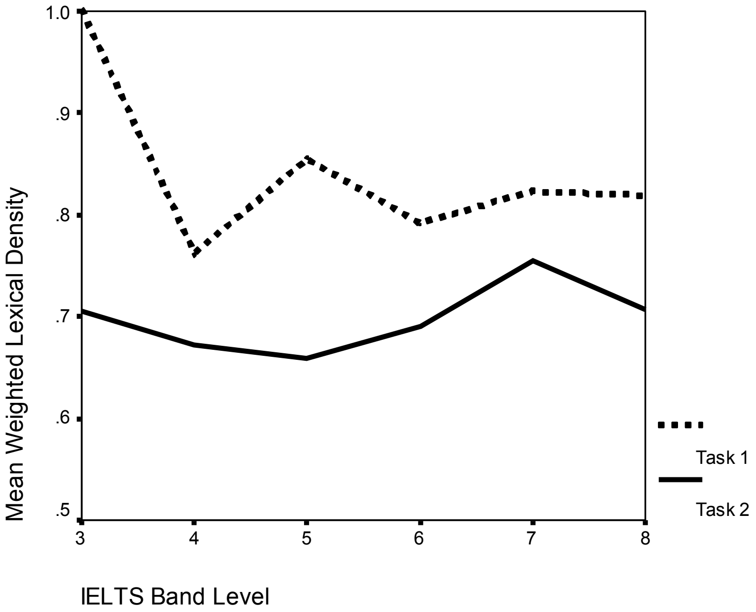


Figure 5.3: Mean lexical density for the whole data set (including standard deviations) by task and IELTS band

Halliday (1985) comments that the complexity of written language lies in the use of more lexical words (in relation to grammatical words) suggesting that as written texts become more complex, they also become more lexical. We therefore expected that the lexical density of the scripts in our data set would rise in line with IELTS band level. We also expected that Task 2 would generate more lexically dense text as it is intended to be the more complex of the two tasks. Figure 5.3 shows that (apart from the data for IELTS band level 3) the mean lexical density of the scripts has a tendency to rise (albeit only slightly) as the IELTS band level increases. These results are in line with our expectations. However, contrary to expectations, Task 1 generates texts with a higher mean lexical density than Task 2.

We performed a one-way ANOVA to check the effect of final band on mean lexical density for Tasks 1 and 2. Table 5.14 shows that there are significant effects for band level on mean lexical density scores for both Task 1 (F = 3.185, p<0.05) and Task 2 (F = 3.990, p<0.05). Post-hoc tests (Tukey HSD) showed that the band levels were significantly different for Task 1 between levels 3 and 4 as well as 3 and 6. For Task 2 the band levels were significantly different between level 7 and levels 4, 5 and 6.

		Sum of Squares	df	Mean Square	F	Sig.
LExWG (1)	Between Groups	.700	5	.140	3.185	.008
	Within Groups	11.817	269	.044		
	Total	12.516	274			
LExWG(2)	Between Groups	.298	5	.060	3.990	.002
	Within Groups	4.012	269	.015		
	Total	4.310	274			

Table 5.14: Summary statistics for the one-way ANOVA band, DV: Mean lexical density for Tasks 1 and 2

We were also interested to see if the relationship between mean lexical density and final band was the same for each L1 group (Table 5.15 and Table 5.16).

	L1 Chinese Mean Lexical Density			L1 Spanish Mean Lexical Density		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 7/0)	0.99 (0.36)	1.57	0.50	-	-	-
Band 4 (N = 29/8)	0.76 (0.18)	1.13	0.53	0.76 (1.17)	0.92	0.46
Band 5 (N = 45/28)	0.91 (0.28)	1.59	0.50	0.78 (0.15)	1.14	0.52
Band 6 (N = 38/38)	0.84 (0.21)	1.48	0.52	0.75 (0.14)	1.07	0.49
Band 7 (N = 9/32)	0.86 (0.21)	1.12	0.55	0.80 (0.14)	1.23	0.57
Band 8 (N = 0/7)	-	-	-	0.79 (0.12)	0.99	0.64

Table 5.15: Mean lexical density for each L1 group (including standard deviations) by IELTS band for Task 1

Table 5.15 shows that the Task 1 mean lexical density for L1 Chinese speakers is not very predictable by IELTS band level but that the Task 1 mean lexical density for L1 Spanish speakers tends to rise (albeit only slightly) in line with IELTS band level. Additionally, the L1 Chinese speakers overall display a higher mean lexical density at every band level than the L1 Spanish speakers. These patterns are illustrated in Figure 5.4.

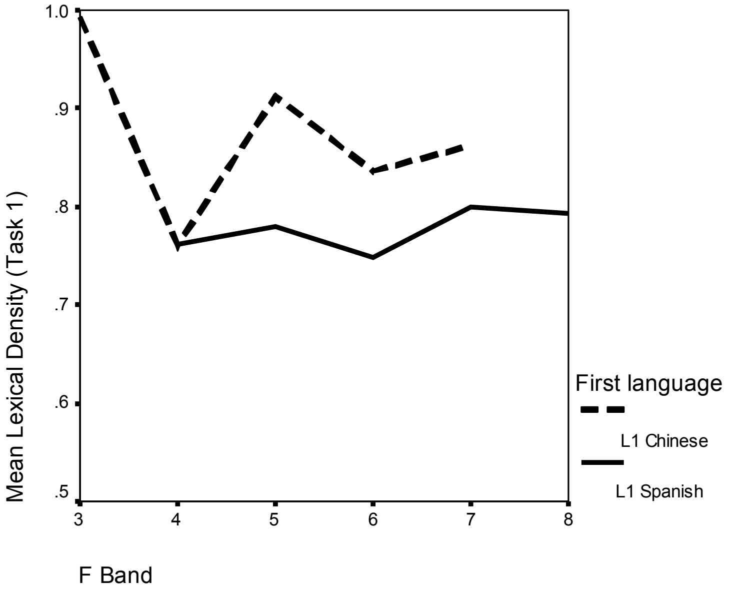


Figure 5.4: Mean lexical density for each L1 group by IELTS band for Task 1

	L1 Chinese Mean Lexical Density			L1 Spanish Mean Lexical Density		
	Mean (SD)	Maximum	Minimum	Mean (SD)	Maximum	Minimum
Band 3 (N = 7/0)	0.69 (0.12)	0.90	0.52	-	-	-
Band 4 (N = 29/8)	0.68 (0.12)	0.99	0.47	0.62 (0.15)	0.80	0.41
Band 5 (N = 45/28)	0.68 (0.09)	0.84	0.49	0.63 (0.12)	0.92	0.42
Band 6 (N = 38/38)	0.70 (0.11)	1.08	0.50	0.68 (0.14)	1.02	0.37
Band 7 (N = 9/32)	0.82 (0.10)	0.95	0.64	0.72 (0.15)	1.15	0.49
Band 8 (N = 0/7)	-	-	-	0.69 (0.06)	0.79	0.63

Table 5.16: Mean lexical density for each L1 group (including standard deviations) by IELTS band for Task 2

Table 5.16 shows that the Task 2 mean lexical density for both L1 groups tends to rise in line with IELTS band level. Once again, the texts written by L1 Chinese speakers tend to have a higher mean lexical density than the L1 Spanish speakers. These patterns are illustrated in Figure 5.5.

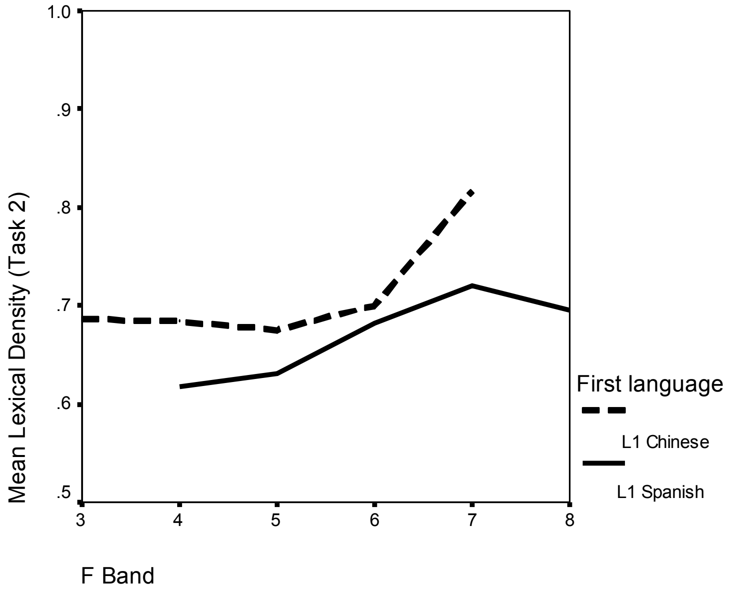


Figure 5.5: Mean lexical density for each L1 group by IELTS band for Task 2

We performed a two-way ANOVA, between groups design, to test the main effects and potential interactions of the factors band level x L1 for both tasks with mean lexical density as the dependent variable, excluding band levels 3 and 8 because we lacked data from one L1 group in each case. Table 5.17 shows that, for Task 1, the main effect of L1 on mean lexical density is significant (F = 4.937, p<0.05), thus confirming the pattern illustrated in Figure 5.4. However, there is no significant effect for IELTS band level and the interaction between the two factors was not significant either.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	.781(a)	7	.112	2.876	.007
Intercept	101.214	1	101.214	2607.000	.000
F_BAND	.185	3	.062	1.590	.193
FIRST_LA	.192	1	.192	4.937	.027
F_BAND * FIRST_LA	.088	3	.029	.755	.520
Error	8.502	219	.039		
Total	159.396	227			
Corrected Total	9.284	226			

a R Squared = .084 (Adjusted R Squared = .055)

Table 5.17: Summary statistics for the two-way ANOVA band x L1, DV: Mean lexical density for Task 1

Table 5.18 shows that, for Task 2, there are significant effects on mean lexical density for both band level (F = 6.697, p<0.01) and L1 (F = 8.243, p<0.05). However, the interaction between the two factors was not significant. Post-hoc tests (Tukey HSD) showed that band levels 4 and 5 were significantly different from 7.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	.326(a)	7	.047	3.143	.003
Intercept	74.091	1	74.091	5006.155	.000
F_BAND	.297	3	.099	6.697	.000
FIRST_LA	.122	1	.122	8.243	.004
F_BAND * FIRST_LA	.035	3	.012	.790	.500
Error	3.241	219	.015		
Total	110.453	227			
Corrected Total	3.567	226			

a R Squared = .091 (Adjusted R Squared = .062)

Table 5.18: Summary statistics for the two-way ANOVA band x L1, DV: Mean lexical density for Task 2

The main findings therefore are:

1. Task 1 generates texts that are higher in mean lexical density than Task 2. This suggests that the nature of the prompt (a table or graph) and the nature of the task (description and evaluation) demand more lexical words (perhaps in the form of adverbs and adjectives) than grammatical words. This merits further investigation through a qualitative analysis of the scripts and the prompts.
2. L1 Chinese speakers on the whole produce more lexically dense text than L1 Spanish speakers.
3. Mean lexical density tends to rise (albeit only slightly) in line with increases in the IELTS band level. This pattern is more marked for Task 2 than for Task 1.

5.5 Lexical sophistication

Measures of lexical sophistication provide insight into the number of unusual or rare words (in relation to specified word lists) that are used by a writer or a group of writers at a particular IELTS band level. We used Range (Nation and Heatley, 1996) to calculate the percentage of words in the scripts at each band level that fell into the first 2000 most frequently used English words (Word Lists 1 and 2), those that were found in the Academic Word List (Word List 3), and those that were not in any of the lists. The results are presented by task and by L1 group. Table 5.19 presents the distribution of words across the lists for both L1 groups in Task 1 and Table 5.20 presents the distribution words across the lists for both L1 groups in Task 2.

	Mean % of Types (L1 Chinese)				Mean % of Types (L1 Spanish)			
	Word List 1	Word List 2	Word List 3	Not in list	Word List 1	Word List 2	Word List 3	Not in list
Band 3	73.60	6.27	6.60	13.53	-	-	-	-
Band 4	68.41	10.47	10.64	10.47	69.31	7.16	11.25	12.28
Band 5	66.85	9.07	12.86	11.23	67.81	8.22	13.93	10.05
Band 6	62.67	9.65	15.00	12.67	64.71	7.48	15.87	11.93
Band 7	65.38	9.58	13.81	11.23	60.21	10.26	17.62	11.92
Band 8	-	-	-	-	64.54	7.57	15.60	12.29

Table 5.19: The % of high and low-frequency words used by both L1 groups at different IELTS band levels for Task 1

	Mean % of Types (L1 Chinese)				Mean % of Types (L1 Spanish)			
	Word List 1	Word List 2	Word List 3	Not in list	Word List 1	Word List 2	Word List 3	Not in list
Band 3	82.31	9.58	3.19	4.91				
Band 4	67.80	12.51	10.57	9.12	72.92	8.50	10.87	7.71
Band 5	60.30	15.39	12.19	12.13	64.37	9.81	14.53	11.28
Band 6	55.15	15.40	14.60	14.85	57.73	10.38	17.59	14.30
Band 7	57.69	13.59	13.91	14.81	56.94	10.85	18.51	13.70
Band 8					70.61	7.27	13.48	8.64

Table 5.20: The % of high and low-frequency words used by both L1 groups at different IELTS band levels for Task 2

Figures 5.6 and 5.7 (see next page) show the relationship between the mean percentage of types from Word Lists 1 and 2 and IELTS band level (for both L1 groups). Both line graphs show a downward trend in the percentage of words taken from Word Lists 1 and 2 suggesting that test-takers at increasingly higher levels of language proficiency will use fewer high-frequency words. It should be noted that there is no effect for L1 and that the pattern of use of high-frequency words is very similar for both the L1 groups. However, it is important to note that:

1. test-takers at lower IELTS band levels use considerably more high-frequency words for Task 2 than for Task 1 but the use of high-frequency words falls away sharply
2. the decline in use of high-frequency words is much more gradual in Task 1 than in Task 2.
3. for each L1 group there is a point at which the use of high-frequency words begins to climb again. The truncated nature of our data set makes it impossible to confirm that this trend continues until IELTS band level 9 but the pattern can be observed for both tasks. For the L1 Chinese group in Task 1, the use of high-frequency words falls away gradually until IELTS band 6 and then begins to rise again. This rise occurs one band later for the L1 Spanish group

(at IELTS band 7). In Task 2 the rise occurs at the same point for each L1 group even though the effect is less marked for the L1 Chinese group.

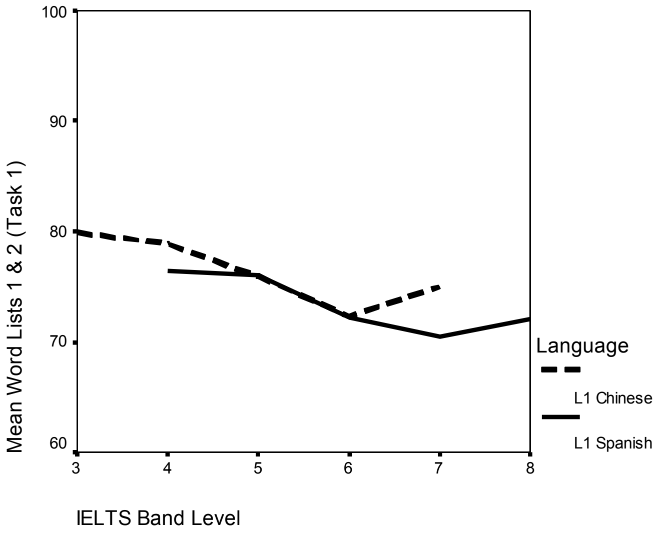


Figure 5.6: Mean percentage of Types from Word Lists 1 and 2 for both L1 groups (represented in Task 1) and IELTS band level

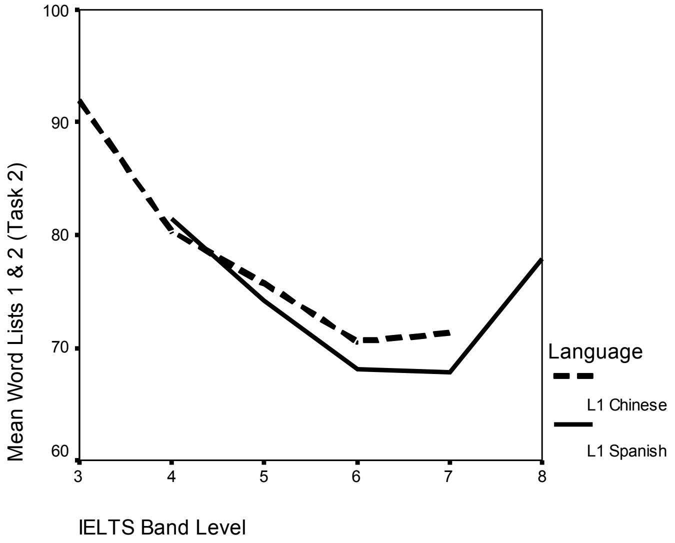


Figure 5.7: Mean percentage of Types from Word Lists 1 and 2 for both L1 groups (represented in Task 2) and IELTS band level

We checked the effect of final band on mean use of Word Lists 1 and 2 by performing a one-way ANOVA. Table 5.21 shows that there are significant effects for band level on mean percentage Word Lists 1 and 2 for Task 2 (F = 22.003, p<0.05) but not for Task 1. Post-hoc tests (Tukey HSD) showed that the band levels were significantly different for Task 2 between band 4 and bands 6 and 7.

		Sum of Squares	df	Mean Square	F	Sig.
WL1_2_1	Between Groups	40.773	3	13.591	4.182	.100
	Within Groups	12.999	4	3.250		
	Total	53.771	7			
WL1_2_2	Between Groups	178.604	3	59.535	22.003	.006
	Within Groups	10.823	4	2.706		
	Total	189.427	7			

Table 5.21: Summary statistics for the one-way ANOVA band, DV: Mean percentage Word Lists 1 and 2 for Tasks 1 and 2

Our analyses therefore indicate that as test-takers' IELTS band scores increase they are more likely to use fewer high-frequency words. However, when test-takers reach a critical IELTS band score the pattern reverses and they appear to draw increasingly on high-frequency vocabulary. It could also indicate a point for that L1 group at which a criterion other than vocabulary becomes more salient to distinguish between one level and another.

5.6 Summary of findings

This analysis of vocabulary richness has investigated the test-takers' vocabulary use from a number of perspectives: lexical output, lexical variation, lexical density, and lexical sophistication. We found the following.

1. Test-takers at higher band levels tend to produce more tokens and types than test-takers at lower band levels. Mean lexical density also rises in line with increases in IELTS band level (though this increase is very gradual) and test-takers are less likely to use high-frequency words as their IELTS band increases.

2. The test-takers' L1 affects the vocabulary richness of their output but in different ways. The L1 Spanish group tended to produce more tokens and types than the L1 Chinese group but in Task 2 the type-token ratio was consistently higher for the L1 Chinese group. The L1 Chinese group also tended to produce more lexically dense text. However, the test-takers' L1 did not affect their sampling of high and low-frequency words (lexical sophistication).

3. The task affects the vocabulary richness of the output in different ways. For instance the type-token ratio for Task 2 was consistently higher than for Task 1 (indicating more lexical variation in the responses for Task 2) but the lexical density (ratio of lexical items to grammatical items) for Task 1 was higher than for Task 2. This indicates that while the test-takers used a greater variety of words for Task 2, these words were not necessarily lexical words. Since we adopted Halliday's approach to lexical density and weighted the low-frequency lexical words more heavily than the high-frequency lexical words it is possible that the lexical words used in Task 2 tended to be high-frequency words. This is in line with the

findings for lexical sophistication which indicate that test-takers (particularly at lower IELTS band levels) were more likely to use high-frequency words in Task 2 than in Task 1.

These findings support the conclusion that cohesive ties in the form of lexical cohesion are more likely at higher IELTS band levels as evinced by greater lexical variation and sophistication in the scripts at these levels. However, the findings also suggest that gains in vocabulary are salient at lower IELTS band levels but other criteria become increasingly salient at higher band levels (perhaps even as early as IELTS band level 7). It would be very interesting to match the findings from this analysis with an investigation into the rating process in order to establish the saliency of different criteria at different IELTS band levels.

6 SYNTACTIC COMPLEXITY

6.1 Review of syntactic complexity measures

Objective measures of syntactic complexity in second language writing owe a lot to work done in assessing levels of first language syntactic maturity, notably the work of Hunt (1965), who demonstrated that sentence length is not a good indicator of language proficiency, and that there is no linear correlation between proficiency and the length of sentences as defined by conventional punctuation. To avoid this obvious discrepancy, Hunt used a different unit of text altogether, which he dubbed the 'minimal terminal unit' – the T-unit – since it is defined as the unit generated when text is divided into the smallest possible independent segments, without leaving sentence fragments behind. Each T-unit consists of a main clause and all the subordinate clauses that belong to it. This preserves both the subordination and the co-ordination of a written text, without yielding an unrealistically high word count per sentence. It effectively redefines the 'sentence' to make it more reliable and suitable for research purposes. This unit, combined with ratio measures of independent and dependent clauses, gives more reliable indications of syntactic maturity in school children's writing.

However, the application of Hunt's T-units to the measurement of language development in several different contexts (Cooper, 1976; Flahive and Snow, 1980) has revealed that T-units are not always ideal for L2 contexts, as they are not sensitive to errors in the text (Flahive and Snow, 1980; Ishikawa, 1995). Wolfe-Quintero et al (1998) comprehensively reviewed the results of studies using 15 different ratio measures for syntactic complexity, and concluded that the best measures overall were *clause per T-unit* and *dependent clause per clause*. However, the authors express some reservations about how well these could discriminate among levels where texts were placed using an holistic rater, rather than by program level, and further difficulties in comparison are presented by the many different definitions of 'clause' that researchers had employed in their studies. Ortega (2003) further cautions against equating complexity in writing to proficiency, noting that writers at the top end of the proficiency scale may well employ rhetorical strategies of complexification at the phrasal, rather than at the clausal, level (nominalisation, for example). She further suggests that the rhetorical styles of the writers' first language are likely to play a significant role in the development of this aspect of language (ibid, pp 514).

This was also the experience of Mayor et al (2000) in their analysis of IELTS scripts by Greek and Chinese L1 writers. They adopted a very rigid definition of a clause, that it should contain a finite verb, in their calculation of clauses per T-unit, but found that this reduced the discriminatory power the measure had for band levels (although it still yielded significant differences for L1 and test version). It seems that by not counting the more sophisticated, non-finite clauses, more proficient writers are penalised. They also note that where there are many errors in language usage, syntactic complexity may be masked, causing quite complex texts to be rated at a lower band.

6.2 Procedure for calculating syntactic complexity

Forty two (42) sample texts were selected to represent the range of levels and L1 in the corpus. These were coded according to the following schema.

Item	**Code**	**Definition**
T-unit	**T-unit**	Independent clause plus all its associated dependent clauses; minimal terminable unit of text.
Main clause	**M-clause**	Clause containing a subject and a predicate which can stand on its own. (predicate = a phrase headed by a TENSED verb)
Dependent clause: relative	**e-rel**	Clauses that have a similar function to an adjective, usually introduced with 'that', 'who', 'which', or clauses that can be replaced by 'this' or 'that'. The verb in the relative clause is TENSED. Relative clauses cannot stand on their own, and can only be present when either embedded in (e-rel) or chained to (c-rel) a main clause.
	c-rel	
Dependent clause: adverbial	**e-adv**	Clauses that have a similar function to adverbs, so their meaning is typically related to location, time, cause, effect, etc. The verb in the adverbial clause is TENSED. Adverbial clauses cannot stand on their own, and can only be present when either embedded in (e-adv) or chained to (c-adv) a main clause.
	c-adv	
Dependent clause: non-finite	**e-nonf**	Clauses that have a verb WITHOUT TENSE (ie, infinitive or participial verb forms). Like other dependent clauses, non-finite clauses cannot stand on their own, but may be either embedded in (e-nonf) or chained to (c-nonf) a main clause.
	c-nonf	
Fragments	**Frag**	Parts of the T-unit that do not contain a verb at all, and are therefore not classified as clauses, but as phrases, of various types.

Table 6.2: Syntactic complexity measures

Where candidates had written a title at the top of their text these were ignored as they were deemed to constitute a different type of text.

In order to capture important information about the sophistication of the writing we decided to use a more complex coding system for clauses, coding for subcategories of dependent clause (adverbial, relative and non-finite). This provided richer data than the overall measures could supply.

We also included ellipsis as part of our calculations because it is an important manifestation of syntactic complexity and enabled us to take into account the ‘embeddedness’ of the dependent clauses. To address the problem of errors masking complexity we applied the strategy of ‘resolving’ problematic clauses before coding (clauses altered in this way were clearly identified with a *). Table 6.3 presents the coding conventions adopted. These codes were used in combination to capture as fully as possible the syntactic complexity of the students’ writing.

Item	Code	Definition
'Double' embedded clause	**Ee**	Dependent clauses that are embedded in an already embedded clause, were given the prefix **ee**- (or **eee**- for those embedded in an **ee**- clause, and so on)
*- clauses	*	Some clauses contained errors that meant that they had to be slightly altered before they could be sensibly coded. These clauses were marked with a * to indicate that the syntactic infelicities (or sometimes punctuation) had been resolved. Records were kept of all these examples.
Ellipsis	**0**	The prefix 0- was added for clauses that were considered to be elliptical. Records were kept of all these examples.

Table 6.3: Coding system adopted for clauses

Clearly, therefore, clauses were tagged at quite a detailed level (for example: e-rel/ ee-rel/ eee-rel). However, to calculate the key measurements of syntactic complexity – *clause per t-unit* and *dependent clause per clause* – detailed codes within categories were combined to give overall figures (such as the number of embedded relative clauses and the number of dependent clauses). Table 6.4 on the following page presents the detailed and combined figures.

6.3 Results

The results of the measures are presented first globally and then separately according to level, L1 and writing task.

	Ordinary	* (Resolved)	0- (Ellipsis)	*0-	Sub-total	Group TOTAL	
T-unit	588	20	24	0	-	632	
M-clause	588	20	24	0	-	632	
e-rel	111	3	47	3	164	185	534 (all dep clauses)
ee-rel	13	0	5	2	20		
eee-rel	1	0	0	0	1		
c-rel	36	0	8	0	-	44	
e-adv	19	0	0	0	19	31	
ee-adv	11	1	0	0	12		
c-adv	84	1	9	0	-	94	
e-nonf	90	0	0	0	90	130	
ee-nonf	34	0	0	0	34		
eee-nonf	4	1	0	0	5		
eeee-nonf	1	0	0	0	1		
c-nonf	48	0	2	0	-	50	
Fragment	322	0	0	0	-	322	

Table 6.4: Total number of categories of clauses found in the 42 samples

clause / t-unit = (632 + 534) / 632 = 1.845
dependent clause / clause = 534 / (632 + 534) = 0.458

Table 6.5 and Figure 6.1 show that the two measures above provide quite different developmental profiles across IELTS band levels. Mean CL/T-unit appears to be a better discriminator of level at lower band levels, but it is not clear that either measure is a good discriminator from level 5 up.

	3	4	5	6	7	8
Mean cl/t-unit	1.432292	3.014958	3.678834	3.586886	4.217046	3.377381
Mean dep cl/cl	0.254233	0.638888	0.784881	0.862093	0.951101	0.809609

Table 6.5: Syntactic complexity measures by level (both language groups)

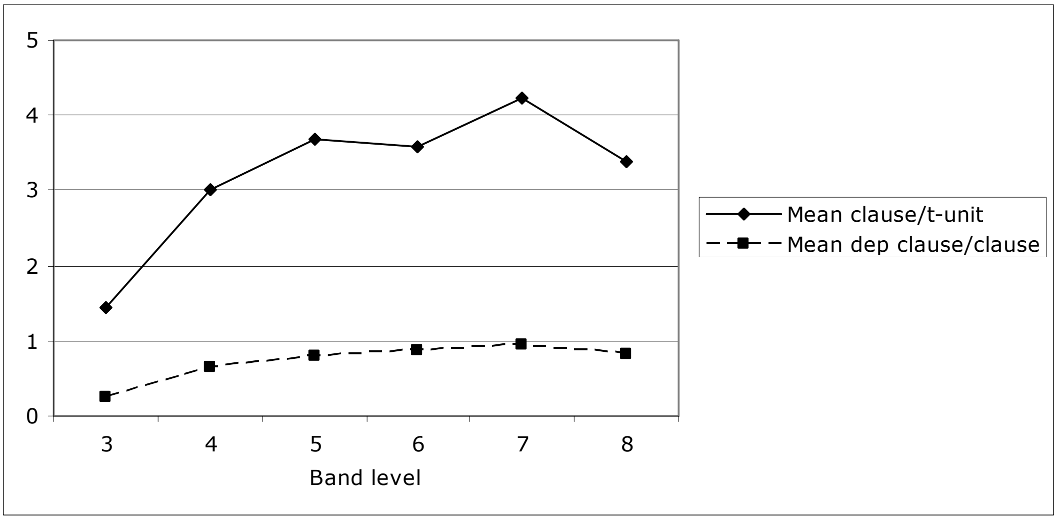


Figure 6.1: Syntactic complexity measures by level (both language groups)

The Writing task appears to have a slight effect on these measures such that Task 2 yields higher scores than Task 1 (cf Table 6.6 and 6.7 and Figure 6.2 and 6.3). However, it is not immediately obvious that this effect is meaningful for our purposes.

	3	4	5	6	7	8
Task 1 Mean c/t-unit	1.43750000	1.36191309	1.52794118	1.65705128	1.51376748	2.06613757
Task 2 Mean c/t-unit	1.42708333	1.65304487	2.15089286	1.92983500	2.70327856	2.43703704

Table 6.6: Syntactic complexity measures by task (both language groups)

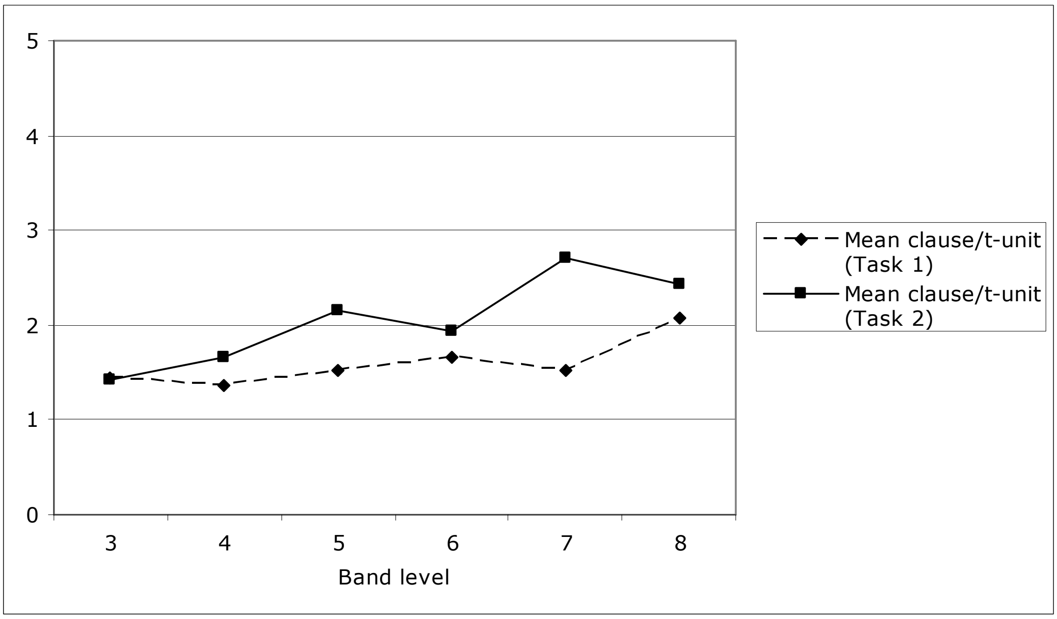


Figure 6.2: Syntactic complexity measures by level (both language groups)

	3	4	5	6	7	8
Task 1 Mean dep. clause /clause	.23333333	.24874666	.27134146	.38728070	.33716692	.50000000
Task 2 Mean dep. clause /clause	.27513228	.39014115	.51353949	.47481246	.61393424	.57947900

Table 6.7: Syntactic complexity measures by task (both language groups)

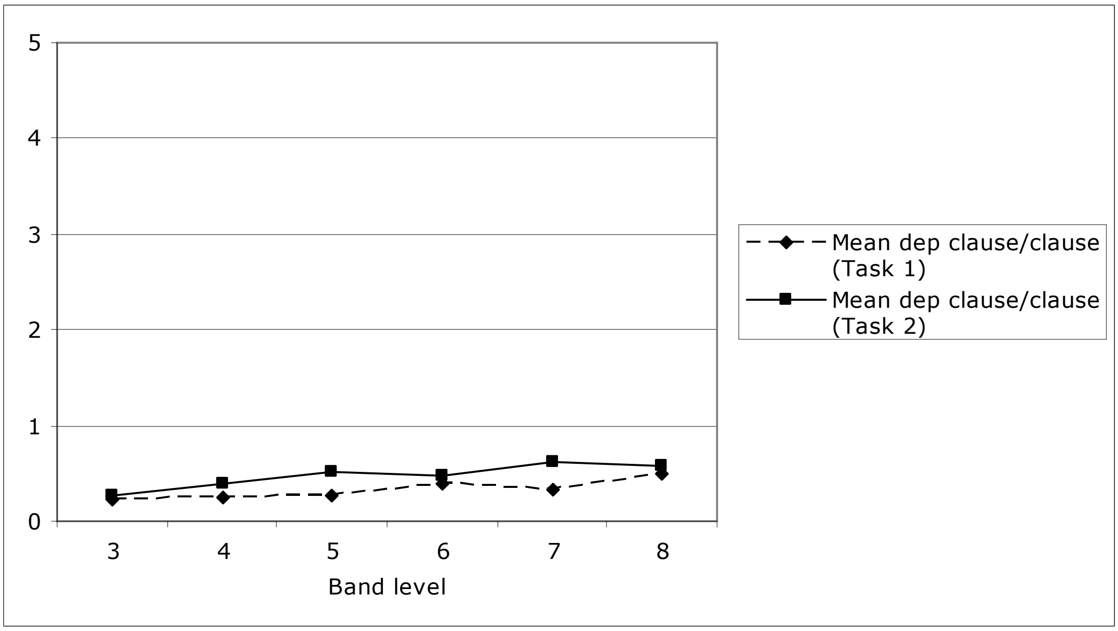


Figure 6.3: Syntactic complexity measures by level (both language groups)

Level	Mean CL / T-UNIT	Mean DEP CL / CL
3	1.43229	0.25423
4	1.46409	0.31045
5	1.40565	0.24949
6	1.72796	0.41465
7	1.8685	0.44014
8	3.52452	0.83872

Table 6.8: Syntactic complexity measures in the L1 Chinese group (by level, on both tasks)

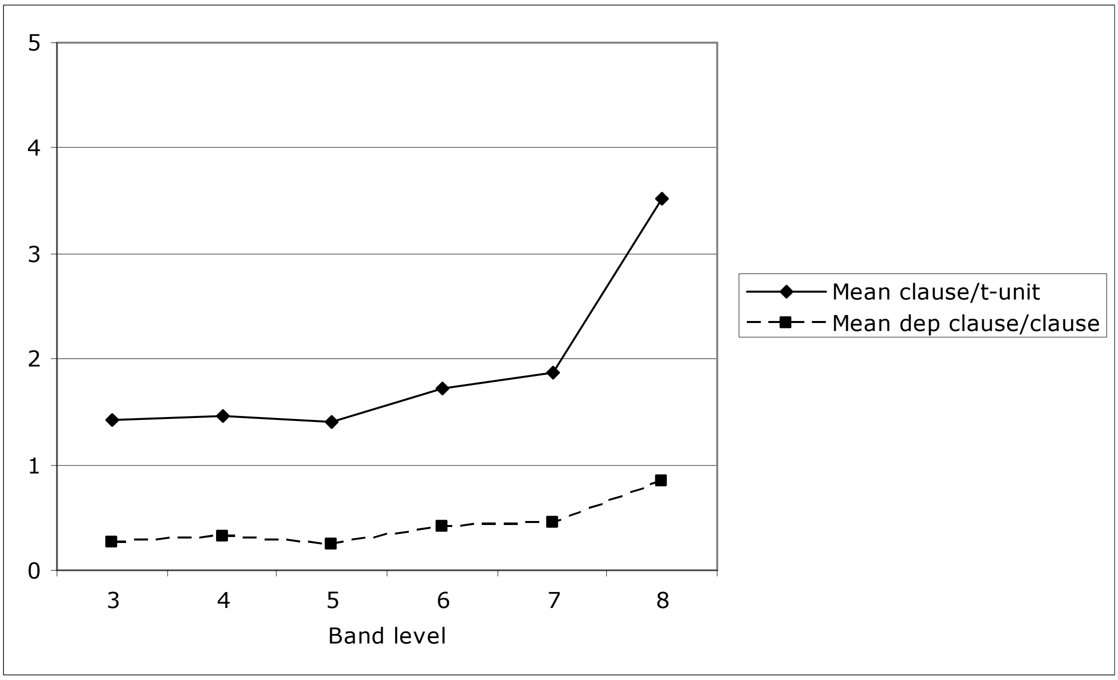


Figure 6.4: Syntactic complexity measures in the L1 Chinese group (by level, on both tasks)

	3	**4**	**5**	**6**	**7**	**8**
Task 1 Mean c/t-unit	1.42708333	1.38811189	1.10000000	1.56410256	1.50480769	2.55555556
Task 2 Mean c/t-unit	1.43750000	1.54006410	1.71130952	1.89181287	2.23219814	2.20000000

Table 6.9: Syntactic complexity measures in the L1 Chinese group (by task)

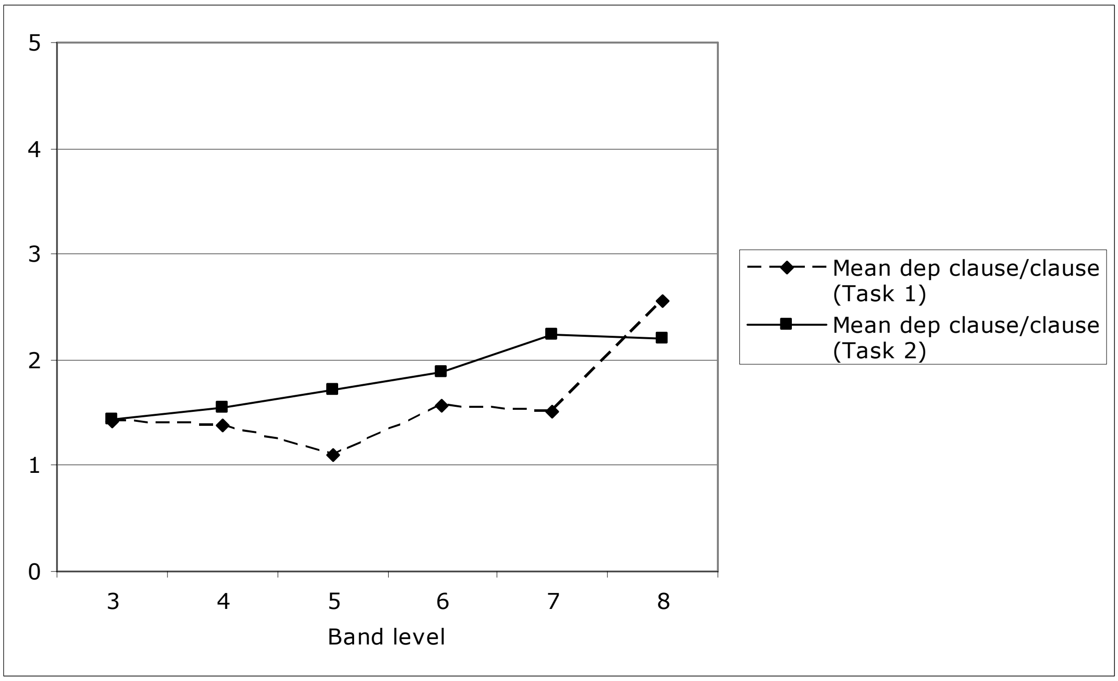


Figure 6.5: Syntactic complexity measures in the L1 Chinese group (by task)

Chinese	3	4	5	6	7	8
Task 1 Mean dep. clause / clause	.23333333	.27022059	.08333333	.35789474	.33119658	.60869565
Task 2 Mean dep. clause / clause	.27513228	.35067568	.41565041	.47140523	.54909318	.54545455

Table 6.10: Syntactic complexity measures in the L1 Chinese group (by task)

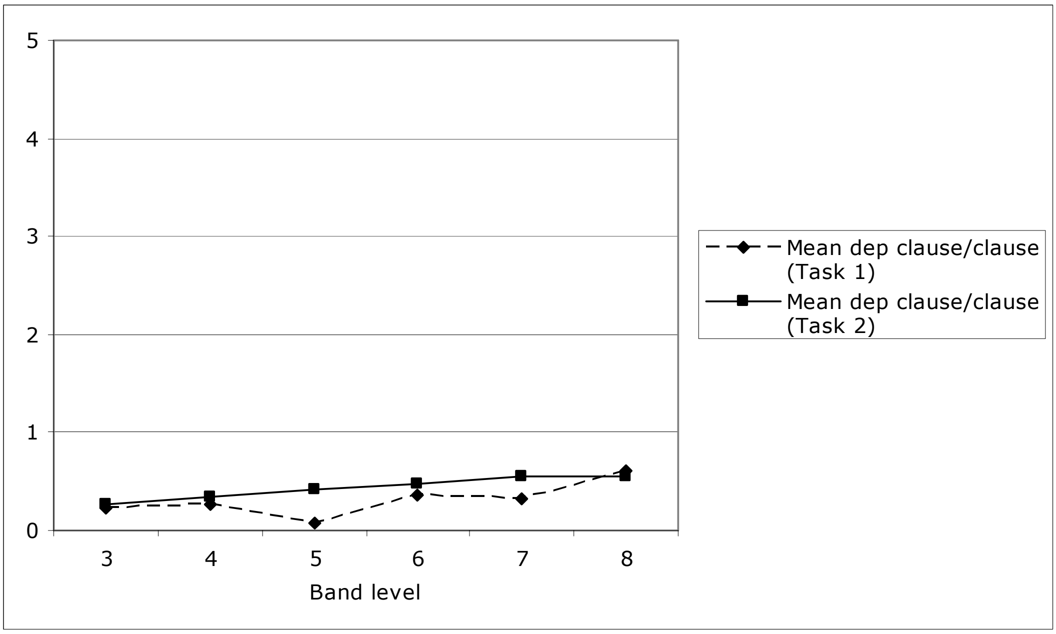


Figure 6.6: Syntactic complexity measures in the L1 Chinese group (by task)

Level	Mean C/T-UNIT	Mean DEP C / CL
4	1.55087	0.32844
5	2.27318	0.53539
6	1.85893	0.44744
7	2.34854	0.51096
8	2.18849	0.52107

Table 6.11: Syntactic complexity measures in the L1 Spanish group (by level, on both tasks)

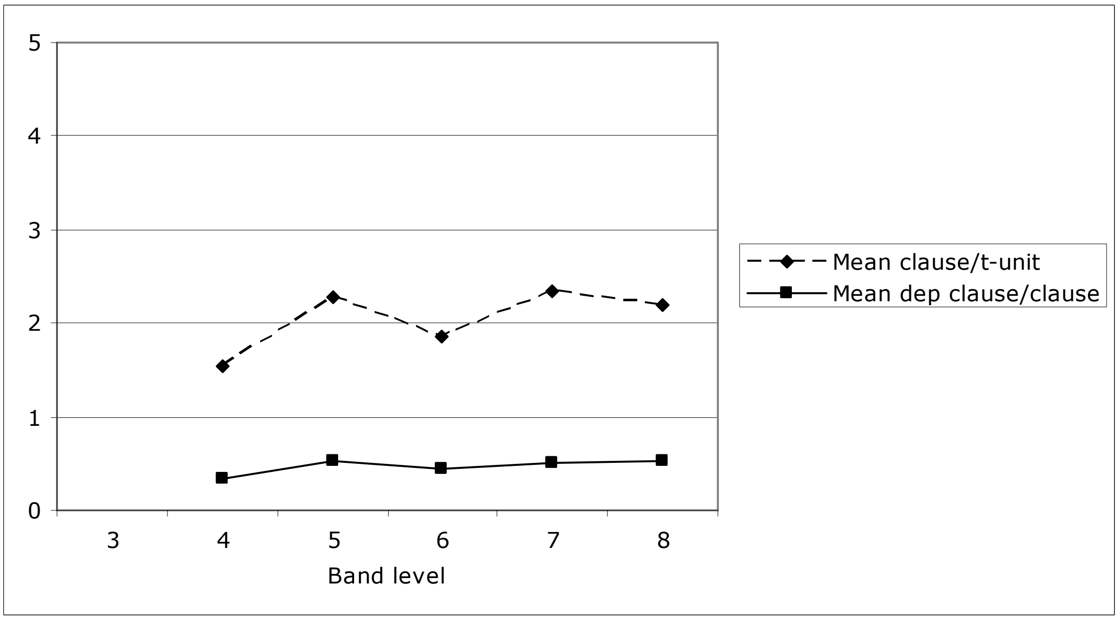


Figure 6.7: Syntactic complexity measures in the L1 Spanish group (by level, on both tasks)

Spanish	**4**	**5**	**6**	**7**	**8**
Task 1 Mean c/t-unit	1.33571429	1.95588235	1.75000000	1.52272727	1.82142857
Task 2 Mean c/t-unit	1.76602564	2.59047619	1.96785714	3.17435897	2.55555556

Table 6.12: Syntactic complexity measures in the L1 Spanish group (by task)

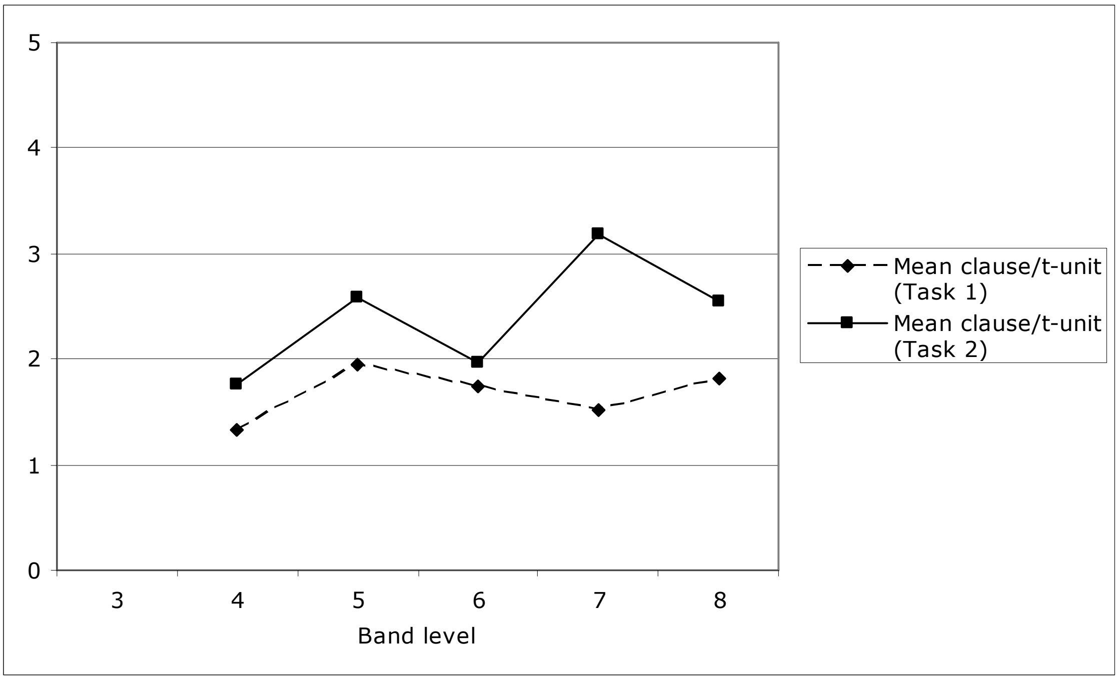


Figure 6.8: Syntactic complexity measures in the L1 Spanish group (by task)

Spanish	**4**	**5**	**6**	**7**	**8**
Task 1 Mean dep. clause /clause	.22727273	.45934959	.41666667	.34313725	.44565217
Task 2 Mean dep. clause /clause	.42960663	.61142857	.47821970	.67877530	.59649123

Table 6.13: Syntactic complexity measures in the L1 Spanish group (by task)

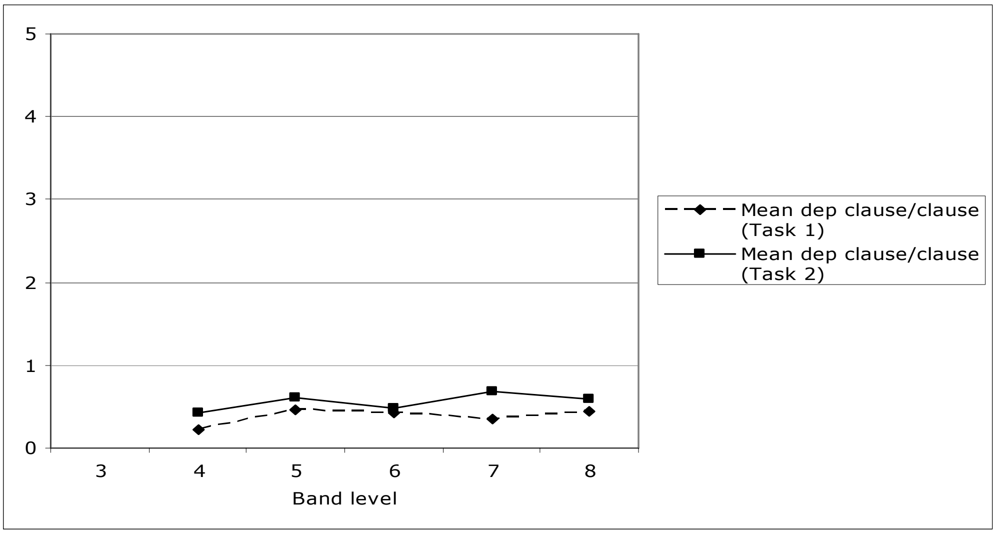


Figure 6.9: Syntactic complexity measures in the L1 Spanish group (by task)

Embedding a dependent clause in the main clause, or in another dependent clause, could be argued to be a more complex way of structuring a sentence than simply chaining the clauses together sequentially. For this reason the coding system was designed to take this into account, to make the identification of embedded clauses possible, as well as clauses that were doubly (and triply) embedded.

In total there were 534 dependent clauses identified in the 42 sample texts coded in this way. Of these, 346 (65%) were embedded in other clauses:

		*	0-	*0-	Sub-total	Group totals
	Ordinary	(Resolved)	(Ellipsis)			
e-rel	111	3	47	3	164	185
ee-rel	13	0	5	2	20	
eee-rel	1	0	0	0	1	
e-adv	19	0	0	0	19	31
ee-adv	11	1	0	0	12	
e-nonf	90	0	0	0	90	130
ee-nonf	34	0	0	0	34	
eee-nonf	4	1	0	0	5	
eeee-nonf	1	0	0	0	1	

Table 6.14: Embeddedness measures (both L1 groups)

Level	e-rel	ee-rel	eee-rel	e-adv	ee-adv	e-nonf	ee-nonf	eee-nonf	eeee-nonf	total
3	9	2	0	0	0	5	0	0	0	16
4	15	2	0	1	1	6	1	0	0	26
5	7	0	0	1	0	6	2	0	0	16
6	15	2	0	1	0	7	2	0	0	27
7	8	0	0	0	0	12	2	1	0	23
8	9	0	0	2	1	6	2	0	0	20
total	63	6	0	5	2	42	9	1	0	**128**

Table 6.15: Embeddedness measures (L1 Chinese group)

Level	e-rel	ee-rel	eee-rel	e-adv	ee-adv	e-nonf	ee-nonf	eee-nonf	eeee-nonf	total
3	-	-	-	-	-	-	-	-	-	-
4	19	1	1	1	1	3	2	1	1	30
5	20	3	0	10	5	20	7	0	0	65
6	22	4	0	0	1	5	2	0	0	34
7	21	4	0	0	2	11	9	3	0	50
8	19	2	0	3	2	9	5	0	0	40
total	101	14	1	14	11	48	25	4	1	**219**

Table 6.16: Embeddedness measures (L1 Spanish group)

These tables show the number of embedded clauses at each level for each language group. No clear trends can be discerned in any category. This may mean that embeddedness is not an indication of syntactic complexity, or that there is no linear relationship between complexity and embeddedness.

Ellipsis was another important factor that was identified during the coding process. A total of 101 elliptical clauses were counted (not including t-units); these were quite evenly distributed among Chinese and Spanish candidates, and no clear trends could be discerned in terms of level.

Level	*0-e-rel	*o-ee-rel	0-c-adv	0-c-nonf	0-c-rel	0-e-adv	0-e-nonf	0-e-rel	0-ee-rel	0-m-clause	total
3	2	2	0	0	0	0	0	4	0	0	**8**
4	0	0	0	0	0	0	0	7	1	3	**11**
5	0	0	2	0	0	0	0	4	0	3	**9**
6	0	0	0	0	1	1	1	3	1	2	**9**
7	0	0	1	1	4	0	0	6	0	2	**14**
8	0	0	2	1	0	0	0	1	0	0	**4**
total	**2**	**2**	**5**	**2**	**5**	**1**	**1**	**25**	**2**	**10**	**54**

Table 6.17: Ellipsis measures (L1 Chinese group)

Level	*0-e-rel	*o-ee-rel	0-c-adv	0-c-nonf	0-c-rel	0-e-adv	0-e-nonf	0-e-rel	0-ee-rel	0-m-clause	total
4	0	0	0	0	0	0	0	4	0	8	**12**
5	0	0	0	0	1	0	0	3	0	2	**6**
6	0	0	1	0	0	0	0	8	1	3	**13**
7	1	0	0	0	1	0	0	5	1	0	**8**
8	0	0	3	0	1	0	0	2	1	1	**8**
total	**1**	**0**	**4**	**0**	**3**	**0**	**0**	**22**	**3**	**14**	**47**

Table 6.18: Ellipsis measures (L1 Spanish group)

The figures above suggest that ellipsis is not an indication of syntactic complexity, or that there is no linear relationship between complexity and ellipsis.

Overall the findings for syntactic complexity measures have not produced a clear developmental picture matching the IELTS band levels 3–8. This could be because syntactic complexity by itself is not a good indicator of increased L2 proficiency as measured by this test or because the specific complexity measures investigated here are not good indicators of increasing IELTS levels.

7 GRAMMATICAL ACCURACY

7.1 Review of measures

The last aspect of task-takers' performance to be investigated in this study is grammatical accuracy. Accuracy measures have been used extensively in research in first and second language development, and they also form part of most rating scales, as indicated in Chapter 2.

The specific measures used here have been borrowed from first and second language acquisition research. They were originally used by Brown (1973) in his seminal longitudinal study of L1 development and soon after adopted by other L1 and L2 researchers (de Villiers 1973, Dulay and Burt 1973, 1974, Bailey, Madden and Krashen 1974, Andersen 1978, Makino, 1980, among many others; see Goldschneider and DeKeyser 2001 for a recent meta-study of this literature).

These studies uncovered a fairly stable set of hierarchies of grammatical accuracy in L2 learners. These hierarchies seemed to be the same regardless of L1 background, type of input received or learning setting (eg instructed vs. naturalistic). For example, Andersen (1978) found the following accuracy hierarchy for verb- and noun-related phenomena:

copula > aspect (ing) > tense (past) > SV agreement (3PS 's')
definite article > plural, indefinite article > possessive 's

That is, the copula was the most accurate verb-related morpheme across L2 learners and an implicational scale could be established such that decreasing levels of accuracy were found in learners as the scale proceeds to the right. The noun-related hierarchy works in the same way.

We decided to investigate a range of morphemes known to be early and late acquired. Our expectation was that the early morphemes (namely copula and plural marking) would perhaps be good discriminators of levels at the low end of the scale and that late morphemes (namely 3rd person singular 's' and passives) may be good discriminators of levels at the higher end of the scale.

A caveat about working with errors is appropriate at this point. The pitfalls of learner error analysis are well-known (see Ellis and Barkhuizen, 2005, for a recent review). Some of the difficulties involved in classifying and quantifying errors have been documented in studies with close links to ours (see for example the discussion in Mayor et al 2002, pp 5 and appendix 1, or Hawkey and Barker, 2004, pp 147-148).

However difficult the task of determining grammatical accuracy may be, we believe that there is enough evidence to suggest that accuracy is a good indicator of L2 proficiency. For example, error rate was found to be a good predictor of proficiency level in Hawkey and Barker (2004, pp 147) and in Wolfe-Quintero et al's meta-study (1998, pp 118). More generally, grammatical accuracy has traditionally been used as a yardstick of development in first and second language acquisition (eg Brown, 1973; de Villiers and de Villiers, 1973; Dulay and Burt, 1973 and 1974; Bailey et al, 1974; Zobl and Liceras, 1994; Goldschneider and DeKeyser, 2001), and the findings of these studies have provided very important insights into the complexities of language development. It is reasonable to expect that the investigation of the development of grammatical accuracy across IELTS band levels will also allow us to shed some light on the research questions at the centre of the present study.

7.2 Procedure for calculating grammatical accuracy

We adopted standard calculations of grammatical accuracy (see Ellis and Barkhuizen, 2005, for more details and critical discussion of this methodology).

Target-Like Use

$$\text{TLU} = \frac{\text{number of correct suppliance in obligatory contexts}}{\text{number of obligatory contexts} + \text{number of suppliance in non-OCs}}$$

7.3 Results

Our findings are compatible with the predictions in the L2 development literature: accuracy on plural and copula was higher than accuracy on SV agreement and passives across levels and L1 groups. SV agreement and passives appear as the best measures of increased proficiency across the whole band range investigated here, and we believe they deserve further investigation, especially third person singular 's' marking, as it was not affected by the learner's L1.

The following tables and graphs summarise the main global findings, followed by more detailed discussion of other findings that should be taken into account when interpreting the global accuracy scores.

	number on 'this'	number on 'that'	number on 'these'	number on 'those'	copula (am, is, are)	copula (was, were)	S-V agreement (3PS 's')	passives
Level 3	0.86	0.92	0.5		0.79	1	0.34	0.41
Level 4	0.873	0.973	0.688	1	0.786	0.661	0.526	0.427
Level 5	0.886	0.983	0.831	0.923	0.895	0.875	0.669	0.659
Level 6	0.982	0.958	0.952	0.909	0.935	0.937	0.81	0.714
Level 7	0.971	1	0.944	1	0.966	0.964	0.838	0.767
Level 8	1		1	1	0.857		1	1

Table 7.1: TLU: L1 Chinese – Tasks 1 and 2

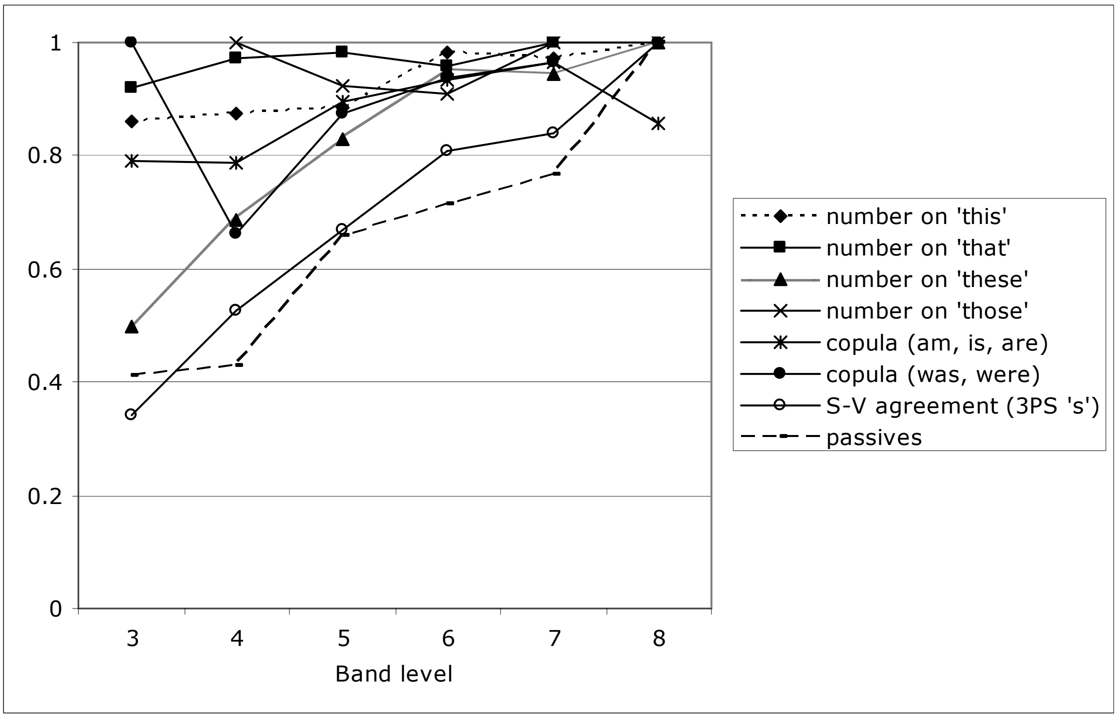


Figure 7.1: TLU: L1 Chinese – Tasks 1 and 2

	number on 'this'	number on 'that'	number on 'these'	number on 'those'	copula (am, is, are)	copula (was, were)	S-V agreement (3PS 's')	Passives
Level 3								
Level 4	0.769	0.944	0.526	0.667	0.946	1	0.348	0.8
Level 5	0.835	0.864	0.313	0.538	0.912	0.914	0.623	0.757
Level 6	0.864	1	0.564	1	0.981	0.9	0.8	0.905
Level 7	0.851	0.969	0.618	0.9	0.993	1	0.913	0.918
Level 8	0.911	1	0.636		0.97	0.889	0.933	0.971

Table 7.2: TLU: L1 Spanish – Tasks 1 and 2

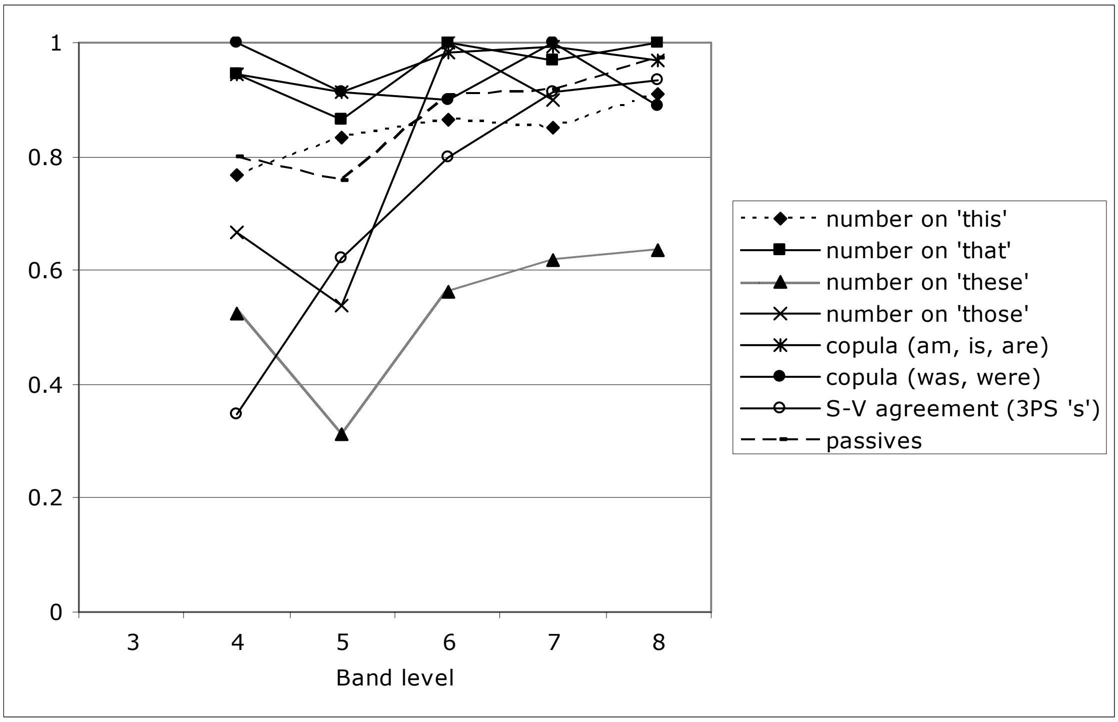


Figure 7.2: TLU: L1 Spanish – Tasks 1 and 2

7.3.1 Default use of the verbs ‘be’ and ‘have’

Even when the meanings assigned to the verbs ‘be’ and ‘have’ were slightly unusual, if the grammatical use in terms of agreement and tense was correct they were computed as correct for purposes of grammatical accuracy in the TLU calculation. However, the use of these two verbs in several instances seemed inappropriate. More specifically, these verbs appear to have been used in a semantic default way (ie they were used in contexts where verbs with more precise meanings could have been used). This is more common in lower level scripts, but there are some cases of high level scripts where this can be seen too.

One implication of this finding may be that vocabulary measures concentrating on range of verb tokens used might be worth looking into in more detail. It may be the case that TTR applied to verbs only will be a good indicator of L2 proficiency, at least when compared with overall TTR or TTR applied to other word categories. This is an empirical matter worth investigating further.

7.3.2 Some number agreement errors seem due to incorrect lexical learning

There are several examples of number agreement errors with certain nouns.

Typical nouns involved:

- information
- news
- people
- women
- police.

Examples:

> and these information are all belong to four countries: Jamaica, Ecuador, Singapore and Bolivia (088-9873-CN002-100104-000-1-6)
>
> To such an extent, the police does not exclude the weapons but require their assistance when living in dangerous environments. (110-3367-CN172-200304-000-2-5)

7.3.3 Prefabricated patterns do not guarantee TL production

Prefabricated patterns are known to be part of development, especially early on. We found that even quite frequent constructions were open to grammatical errors.

> It is well know to everybody that the socity need competition.
> (037-1997-CN902-230202-083-2-4)
>
> It can be clear seen that carbon dioxide produced from power stations takes the biggest amount all over the 2 decades. (025-4749-CN911-121002-090-1-6)

7.3.4 Difficulties determining obligatory contexts with low-level scripts

Sometimes the clause structure of sections of texts is very hard to analyse. This is more frequent at the lowest levels of proficiency. When agreement could not be reached on what the correct analysis of an item should be, the item was discarded from the accuracy analysis. A log was kept of which items were discarded, so further analyses of these contexts could be done in future if required.

7.3.5 Difficulty distinguishing formulaic vs. productive use of language

We treated all language produced by the learners as productive language, as it was felt that decisions about whether specific utterances were cases of formulaic or productive use were to a large degree arbitrary when based on a reader’s judgement. More careful analysis of formulaic and/or repetitive language use seems to us to be an interesting area for further exploration, but appropriate methodological techniques need to be developed before this can be done in a reliable way.

We felt that this was beyond what could be realistically achieved in the time available. However, we have labelled the cases that at first sight seem obvious candidates to be classified as formulaic use in the TLU analysis sheets to facilitate further analysis.

7.3.6 Inflation of scores by repetition of certain structures

In some cases, writers have used the same lexical or grammatical structure repeatedly and this may be interpreted as an inflation factor. It is difficult to decide whether these repeated structures should be discarded. We decided to leave them in, but would like the reader to be aware of this.

Future studies could also look in more detail at the error clusters identified. This work could in turn be used to build materials for teachers, course directors, testers and other stake-holders. For example, concrete examples of the language described in marking guidelines could be identified from the coded database.

8 CONCLUSIONS

The principle objectives of this study have been to document the linguistic markers of different levels of English language writing proficiency defined by the academic version of the IELTS writing module. We sampled 275 scripts from test-takers in two major L1 groups (L1 Chinese and L1 Spanish) at levels 3–8 on the IELTS band scale. We analysed the use of the demonstratives ‘this’, ‘that’, ‘these’ in our corpus, finding that L1 interacts with the task to affect demonstrative use in a number of ways.

1. L1 Spanish speakers use approximately 50% more demonstratives than L1 Chinese speakers.
2. For L1 Chinese speakers the task affects the number of demonstratives used but the relationship between demonstrative use and IELTS band level remains the same. For L1 Spanish speakers, the number of demonstratives used is fairly stable but the relationship between demonstrative use and IELTS band level differs from Task 1 to Task 2.

We observed, however, that use of demonstratives appears to tail off at higher levels of language proficiency, suggesting that other cohesive ties (such as lexical ties) come into use. We would therefore expect performances at higher IELTS band levels to display greater lexical variation and sophistication.

The findings from our analysis of vocabulary richness support this expectation in the sense that scripts at increasing IELTS band levels displayed greater lexical variation and sophistication. Other findings were:

1. The L1 of the test-taker affects lexical output, lexical variation and lexical density but it does not affect lexical sophistication.
2. The task affects vocabulary richness in different ways. Task 1 scripts tend to be more lexically dense than Task 2 scripts and also appear to generate the use of fewer high-frequency words as a proportion of total words. However, Task 2 scripts are more lexically varied (as measured by type-token ratio).

Our results also suggest that gains in vocabulary are salient at lower IELTS band levels but that other criteria become increasingly salient at higher band levels (perhaps even as early as IELTS band level 7). It would be very interesting to take these findings forward by matching these analyses with an investigation into the rating process in order to establish the saliency of different criteria at different IELTS band levels. Secondly, future research could explore ways of modelling the interactions between the different measures in the expectation that different measures group together to contribute to test-takers’ scores at different IELTS band levels.

Our findings for complexity measures were somewhat disappointing, in that by themselves none of the measures investigated seemed to provide a good predictor of IELTS band level. However, negative findings are also to some extent useful findings in that they should help future researchers to decide which measures are not worth pursuing for tracking increasing levels of proficiency.

Finally, the analysis of grammatical accuracy proved to be quite informative, and the predictions from the literature on L2 development were largely confirmed by our data. This suggests to us that future research on predictors of levels of L2 proficiency as measured by the IELTS academic writing tasks should look further into the accuracy of grammatical areas such as SV agreement and passives, as these proved good discriminators of level regardless of L1 and writing task.

REFERENCES

Andersen, RW, 1978, ‘An implicational model for second language research’ in *Language Learning,* vol 28, pp 221-282

Bachman, LF and Cohen, AD (eds), 1998, *Interfaces between second language acquisition and language testing research*, Cambridge University Press, Cambridge

Bailey, N, Madden, CG and Krashen, SD, 1974, ‘Is there a ‘natural sequence’ in adult second language learning?’ in *Language Learning,* vol 24, pp 235-243

Botley, SP, 2000, Corpora and discourse anaphora: using corpus evidence to test theoretical aims, PhD thesis, Lancaster University

Botley, S and McEnery, AM, 2000, ‘Discourse anaphora: the need for synthesis’ in *Corpus-based and computational approaches to discourse anaphora,* eds S Botley and AM McEnery, John Benjamins, Amsterdam, pp 1-41

Brown, R, 1973, *A first language: the early stages*, Harvard University Press, Cambridge, MA

Church, KW and Gale, WA, 1995, ‘Poisson mixtures’ in *Ị atural Language Engineering,* vol 1(2), pp 163-190

Clahsen, H and Muysken, P, 1986, ‘The availability of universal grammar to adult and child learners: a study of the acquisition of German word order’ in *Second Language Research,* vol 2, pp 93-119

Clahsen, H and Muysken, P, 1989, ‘The UG paradox in SLA’ in *Second Language Research,* vol 5, pp 1-29

Cooper, TC, 1976, ‘Measuring written syntactic patterns of second language learners of German’ in *Journal of Educational Research,* vol 69, pp 176-183

Council of Europe, 2001, Common European framework of reference: learning, teaching, assessment, Cambridge University Press, Cambridge

Coxhead, A, 2000, ‘A new academic word list’ in *TESOL Quarterly,* vol 34, pp 213-238

De Villiers, J, and de Villiers, P, 1973, ‘A cross-sectional study of the development of grammatical morphemes in child speech’ in *Journal of Psycholinguistic Research,* vol 2, pp 267-278

Douglas, D, 2001, ‘Performance and consistency in second language acquisition and language testing research: a conceptual gap’ in *Second Language Research,* vol 17, pp 442-456

Dulay, H and Burt, M, 1973, ‘Should we teach children syntax?’ in *Language Learning,* vol 23, pp 245-258

Dulay, H and Burt, M, 1974, ‘Natural sequences in child second language acquisition’ in *Language Learning,* vol 24, pp 37-53

Durán, P, Malvern, D, Richards, B and Chipere, N, 2004, ‘Developmental trends in lexical diversity’ in *Applied Linguistics,* vol 25(2), pp 220-242

Ellis, R, 2001, ‘Some thoughts on testing grammar: An SLA perspective’ in *Experimenting with uncertainty: essays in honour of Alan Davies*, eds C Elder, A Brown, E Grove, K Hill, N Iwashita, T Lumley, T McNamara, and K O’Loughlin, University of Cambridge Local Examinations Syndicate (UCLES), Cambridge, pp 251 – 263

Ellis, R and Barkhuizen, G, 2005, *Analysing learner language*, Oxford University Press, Oxford

Engber, C, 1995, 'The relationship of lexical proficiency to the quality of ESL compositions' in *Journal of Second Language Writing,* vol 4, pp 139-155

Flahive, DE and Snow, BG, 1980, 'Measures of syntactic complexity in evaluating ESL compositions' in *Research in language testing*, eds JW Oller and K Perkins, Newbury House, Rowley, MA, pp 171-176

Flowerdew, L, 1998, 'Integrating expert and interlanguage computer corpora findings on causality: discoveries for teachers and students' in *English for Specific Purposes,* vol 17, pp 329-345.

Ghazzoul, N, in progress, 'Coherence in English academic writing of Arab EFL learners with special reference to Syrian and Emirati university students', PhD thesis in progress, Lancaster University

Goldschneider, JM and DeKeyser, RM, 2001, 'Explaining the 'natural order of L2 morpheme acquisition' in English: a meta-analysis of multiple determinants' in *Language Learning,* vol 51, pp 1-50

Halliday, MAK, 1985, *Introduction to functional grammar,* Arnold, London

Halliday, MAK, 1994, *An introduction to functional grammar* (2nd Edition), Edward Arnold, London

Halliday, MAK and Hasan, R, 1976, *Cohesion in English*, Longman Group Ltd, London

Hawkey, R and Barker, F, 2004, 'Developing a common scale for the assessment of writing' in *Assessing Writing,* vol 9, pp 122-159

Homburg, TJ, 1984, 'Holistic evaluation of ESL compositions: can it be validated objectively?' in *TESOL Quarterly,* vol 18, pp 87-107

Hunt, KW, 1965, *Grammatical structures written at three grade levels*, The National Council of Teachers of English, Urbana, IL

Hyltenstam, K and Pienemann, M, 1985, *Modelling and assessing second language development*, Multilingual Matters, Clevedon

International English Language Testing System (IELTS), <http://www.ielts.org> (accessed 09 January 2006)

International English Language Testing System (IELTS), 2005, *IELTS Handbook*, <http://www.ielts.org/_lib/pdf/1649_IELTShbk_2005.pdf> (accessed 14 September 2006)

Ishikawa, S, 1995, 'Objective measurement of low proficiency EFL narrative writing' in *Journal of Second Language Writing,* vol 4, pp 51-70

Kennedy, C and Thorp, D, 2002, A corpus investigation of linguistic responses to an IELTS Academic Writing task, IELTS British Council Research Programme

Larsen-Freeman, D, 1978, 'An ESL index of development' in *TESOL Quarterly,* vol 12, pp 439-448

Laufer, B, 2001, 'Quantitative evaluation of vocabulary' in *Experimenting with uncertainty: essays in honour of Alan Davies*, eds C Elder, A Brown, E Grove, K Hill, N Iwashita, T Lumley, T McNamara, and K O'Loughlin, University of Cambridge Local Examinations Syndicate (UCLES), Cambridge, pp 241-250

Laufer, B and Nation, P, 1995, 'Vocabulary size and use: lexical richness in L2 written production' in *Applied Linguistics,* vol 16, pp 307-322

Makino, T, 1980, 'Acquisition order of grammatical morphemes by Japanese secondary school students' in *Journal of Hokkaido University of Education,* vol 30, pp 101-148

Malvern, D and Richards, B, 2002, 'Investigating accommodation in language proficiency interviews using a new measure or lexical diversity' in *Language Testing,* vol 19, pp 85-104

Mayor, B, Hewings, A, North, S, Swann, J and Coffin, C, 2002, *A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2*, IELTS British Council Research Programme

McNamara, T, 1996, Measuring second language performance, Longman, London

Meara, P and Miralpeix, I, 2004, *D_Tools*, Lognostics (Centre for Applied Language Studies, University of Wales Swansea), Swansea

Meisel, JM, 1997, 'The acquisition of the syntax of negation in French and German: contrasting first and second language development' in *Second Language Research,* vol 13, pp 227-263

Muhr, T, 2005, *Atlas-ti*, <http://www.atlasti.com/> (accessed 14 September 2006)

Nation, P and Heatley, A, 1996, *Range*, School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington

O'Loughlin, K, 2001, *The equivalence of semi-direct speaking tests*, University of Cambridge Local Examinations Syndicate and Cambridge University Press, Cambridge

Odlin, T, 2003, 'Cross-linguistic influence' in *Handbook of Second Language Acquisition*, eds CJ Doughty and MH Long, Blackwell, Malden, MA, pp 436-486

Ortega, L, 2003, 'Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing' in *Applied Linguistics,* vol 24, pp 492-518

Perdue, C and Klein, W, 1993, 'Concluding remarks', in *Adult language acquisition: Crosslinguistic perspectives. Volume II: The results*, ed C Perdue, Cambridge University Press, Cambridge, pp 253-272

Read, J, 2000, *Assessing vocabulary*, Cambridge University Press, Cambridge

Read, J, 2005, 'Applying lexical statistics to the IELTS speaking test', *Cambridge Research Ị otes,* vol 20, pp 12-16

Shaw, SD, 2002, 'IELTS writing: revising assessment criteria and scales (Phase 2)', *Cambridge Research Ị otes,* vol 10, pp 10-13

Shaw, SD, 2004, 'IELTS writing: revising assessment criteria and scales (Phase 3)', *Cambridge Research Ị otes,* vol 16, pp 3-7

Shohamy, E, 1998, 'How can language testing and SLA benefit from each other? The case of discourse' in *Interfaces between second language acquisition and language testing research*, eds LF Bachman and AD Cohen, 1998, Cambridge University Press, Cambridge, pp 156-176

Skehan, P, 1989, Individual differences in second language learning, Arnold, London

Slavoff, GR and Johnson, J, 1995, 'The effects of age on the rate of learning a second language' in *Studies in Second Language Acquisition,* vol 17, pp 1-16

Teddick, D, 1990, 'ESL writing assessment: subject matter knowledge and its impact on performance' in *English for Specific Purposes,* vol 9, pp 123-143

Ure, J, 1971, 'Lexical density and register differentiation' in *Applications of Linguistics*, eds GE Perren and JLM Trimm, Cambridge University Press, Cambridge

Weigle, SC, 2002, *Assessing writing*, Cambridge University Press, Cambridge

West, M, 1953, A general service list of English words, Longman, London

Wolfe Quintero, K, Inagaki, S and Kim, H-Y, 1998, *Second language development in writing: measures of fluency, accuracy and complexity*, in Technical Report 17. Honolulu: University of Hawai'i at Manoa, Second Language Teaching and Curriculum Centre

Zobl, H and Liceras, JM, 1994, 'Review article: functional categories and acquisition orders' in *Language Learning,* vol 44, pp 159-180

APPENDIX 1

The mean frequency of use of the demonstratives 'this', 'that', 'these' and 'those' (including standard deviations) according to L1 and IELTS band level for Task 1

	L1 Chinese Means (SD)				L1 Spanish Means (SD)			
	this	that	these	those	this	that	these	those
Band 3 (N = 7/0)	0.14 (0.38)	0.71 (1.11)	0.14 (0.38)	0.00 (0.00)	-	-	-	-
Band 4 (N = 29/8)	0.66 (0.94)	0.38 (0.49)	0.14 (0.44)	0.03 (0.19)	2.13 (1.89)	1.38 (2.77)	0.87 (2.10)	0.25 (0.46)
Band 5 (N = 45/28)	0.67 (0.80)	0.16 (0.42)	0.76 (1.13)	0.09 (0.36)	2.21 (1.99)	0.57 (0.96)	0.29 (0.98)	0.07 (0.26)
Band 6 (N = 38/38)	0.76 (1.05)	0.79 (1.17)	0.53 (0.69)	0.13 (0.48)	2.11 (1.57)	0.29 (0.46)	0.39 (0.86)	0.11 (0.39)
Band 7 (N = 9/32)	0.67 (1.00)	1.44 (1.94)	0.44 (0.73)	0.22 (0.44)	1.69 (1.45)	0.50 (0.84)	0.47 (0.72)	0.19 (0.47)
Band 8 (N = 0/7)	-	-	-	-	2.57 (2.15)	0.43 (0.79)	1.00 (1.00)	0.00 (0.00)

The mean frequency of use of the demonstratives 'this', 'that', 'these' and 'those' (including standard deviations) according to L1 and IELTS band level for Task 2

	L1 Chinese Means (SD)				L1 Spanish Means (SD)			
	this	that	these	those	this	that	these	those
Band 3 (N = 7/0)	1.14 (2.61)	0.29 (0.49)	0.00 (0.00)	0.00 (0.00)	-	-	-	-
Band 4 (N = 29/8)	1.62 (1.61)	1.21 (1.63)	0.55 (0.78)	0.03 (0.19)	1.38 (1.12)	1.25 (1.04)	0.25 (0.71)	0.00 (0.00)
Band 5 (N = 45/28)	1.33 (1.35)	0.89 (1.27)	0.44 (0.87)	0.09 (0.29)	1.86 (1.74)	1.25 (1.43)	0.29 (0.66)	0.25 (0.52)
Band 6 (N = 38/38)	1.71 (1.37)	1.03 (1.50)	0.71 (1.18)	0.37 (0.79)	2.66 (2.18)	1.03 (0.94)	0.53 (1.03)	0.32 (0.62)
Band 7 (N = 9/32)	1.44 (1.33)	0.33 (0.50)	0.56 (0.53)	0.89 (1.05)	2.56 (1.90)	0.63 (0.97)	0.53 (0.80)	0.22 (0.42)
Band 8 (N = 0/7)	-	-	-	-	4.00 (3.11)	0.14 (0.38)	0.29 (0.49)	0.00 (0.00)

APPENDIX 2: 50 MOST FREQUENT WORDS

The 50 most frequent words in the L1 Chinese scripts

N	Word	Freq.	%	N	Word	Freq.	%
1	THE	5,093	6.83	26	ON	368	0.49
2	AND	2,248	3.01	27	I	361	0.48
3	IN	2,192	2.94	28	THIS	360	0.48
4	OF	1,920	2.57	29	SOME	345	0.46
5	TO	1,866	2.50	30	THAN	342	0.46
6	IS	1,435	1.92	31	MEN	340	0.46
7	A	952	1.28	32	BUT	338	0.45
8	THAT	789	1.06	33	WILL	334	0.45
9	CAN	751	1.01	34	COUNTRIES	310	0.42
10	IT	745	1.00	35	WHICH	310	0.42
11	ARE	735	0.99	36	DO	293	0.39
12	MORE	680	0.91	37	THERE	293	0.39
13	WOMEN	673	0.90	38	SO	285	0.38
14	FOR	656	0.88	39	ABOUT	271	0.36
15	PEOPLE	595	0.80	40	RATE	263	0.35
16	THEY	546	0.73	41	HAS	257	0.34
17	FROM	538	0.72	42	SHOULD	257	0.34
18	AS	501	0.67	43	POPULATION	255	0.34
19	WE	492	0.66	44	BY	252	0.34
20	HAVE	446	0.60	45	ALL	235	0.32
21	BE	445	0.60	46	MILLION	233	0.31
22	WITH	391	0.52	47	LITERACY	231	0.31
23	THEIR	378	0.51	48	OR	229	0.31
24	NOT	370	0.50	49	ONLY	226	0.30
25	S	369	0.49	50	FEMALE	225	0.30

The 50 most frequent words in the L1 Spanish scripts

N	Word	Freq.	%	N	Word	Freq.	%
1	THE	4,092	6.79	26	MORE	304	0.50
2	IN	2,098	3.48	27	BUT	297	0.49
3	OF	2,005	3.33	28	LANGUAGE	297	0.49
4	AND	1,847	3.06	29	NOT	295	0.49
5	TO	1,636	2.71	30	HAS	285	0.47
6	A	1,325	2.20	31	THEIR	285	0.47
7	IS	1,122	1.86	32	WORLD	283	0.47
8	THAT	939	1.56	33	I	273	0.45
9	ARE	573	0.95	34	OR	271	0.45
10	FOR	560	0.93	35	OTHER	251	0.42
11	HAVE	556	0.92	36	ON	230	0.38
12	IT	553	0.92	37	ALL	229	0.38
13	THIS	544	0.90	38	BY	217	0.36
14	AS	536	0.89	39	MOBILE	207	0.34
15	WITH	534	0.89	40	COUNTRY	205	0.34
16	COUNTRIES	517	0.86	41	THERE	199	0.33
17	PEOPLE	507	0.84	42	BECAUSE	196	0.33
18	WE	428	0.71	43	WORKFORCE	196	0.33
19	BE	427	0.71	44	AN	195	0.32
20	WOMEN	364	0.60	45	THAN	194	0.32
21	ENGLISH	333	0.55	46	LANDLINE	185	0.31
22	THEY	333	0.55	47	IMPORTANT	178	0.30
23	EDUCATION	330	0.55	48	ONLY	177	0.29
24	PHONES	324	0.54	49	ONE	176	0.29
25	CAN	310	0.51	50	FROM	170	0.28