

# Analysing written language

Day Two

14.00 – 17.30



Qualitative Methods

[www.lsa.umich.edu/eli](http://www.lsa.umich.edu/eli)

- Why?
- Measures
  - Cohesive Devices
  - Vocabulary Richness
  - Syntactic Complexity
  - Grammatical Accuracy
- Qualitative or Quantitative?
- Working with data
- A practical application



# Why?

- Investigation of input texts (providing an empirical basis for difficulty claims and perhaps ensuring equivalence across test administrations)
- Validation of rating scales (providing an empirical basis for descriptions of competence at each performance level)
- Providing teachers with a description of language development that can be used for diagnosis, agenda setting and curriculum planning.



# Measures

- Vocabulary Richness
  - Lexical output
    - Preliminaries
      - Tokens | Types | Lemmas
      - She **begged** **for** forgiveness, **begging** also **for** mercy.
    - Total number of tokens – total number of words
    - Total number of types – total number of different word forms (where tokens have been lemmatised)



- Vocabulary Richness
  - Lexical variation/diversity
    - Standardised Type-Token Ratio (TTR)
      - $(\text{Types} \div \text{Tokens}) \times 100$
    - D-value (Malvern & Richards, 2002)
      - Multiple samples from the text (of increasingly larger chunks)
      - TTR calculations for each sample
      - Graphs plotting the TTR calculations
      - D-value represents the fit between the actual curve obtained and the curve expected from mathematical models



- Vocabulary Richness
  - Lexical density
    - Preliminaries
      - Lexical/content words e.g. verbs, adjectives, adverbs, nouns
      - Grammatical words e.g. prepositions, conjunctions
    - Taking account of word frequency (O'Loughlin, 2001)
    - $LD = [(High\ frequency\ lexical\ words \div 2) \times Low\ frequency\ lexical\ words] \div Grammatical\ words$



- Vocabulary Richness
  - Summary of measures so far:
    - Number of words produced (lexical output)
    - Ratio of different words in a text (lexical variation/diversity)
    - Ratio of content (lexical) words in a text (lexical density)
  - NOTE:
    - Lexical error unaccounted for
    - No insight into the number of unusual or rare words used (lexical sophistication)
    - No exploration of the use of multi-word lexical structures (formulaic sequences)



- Vocabulary Richness
  - Error-free lexical variation
    - Suggested by Engber (1995)
    - Calculates the % of lexical errors in a text
    - Criticisms:
      - It does not distinguish between errors in types and tokens and could result in double-counting of errors (Laufer & Nation, 1995).
      - It is not always easy to distinguish between lexical and grammatical errors.
      - The framework does not take into account the relative seriousness of different errors (Read, 2000).



- Vocabulary Richness
  - Lexical sophistication
    - Preliminaries
      - 1000 and 2000 word lists (West, 1953)
      - Academic word list (Coxhead, 2000)
    - All the words in a script are classified into four categories
      - 1<sup>st</sup> 1000 most frequently occurring words
      - 2<sup>nd</sup> 1000 most frequently occurring words
      - Academic word list
      - Words not contained in the first three lists



- Vocabulary Richness
  - Formulaic sequences
    - “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored or retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (Wray, 2002: 9)
    - Academic Formula List (Ellis et al., 2008)
    - Ohlrogge (2009)



- Syntactic Complexity

- Clause per T-unit

- The 'T-unit' (minimal terminal unit) - 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.

She begged for forgiveness, **begging also for mercy.**



- Syntactic Complexity
  - Dependent clause per clause
    - Count the number of clauses in each text
    - Count the number of dependent clauses.
    - Divide the number of dependent clauses by the number of clauses.

She begged for forgiveness **and mercy**.

She begged for forgiveness **and asked for money**.



- **Grammatical Accuracy**
  - You need to specifically choose the grammatical features that you want to check.
    - number on demonstratives (this, that, these, those)
    - copula in the present and past tense (e.g. verb ‘to be’)
    - subject-verb agreement on main verbs in the present (3rd person singular ‘s’)
    - passives.



- Grammatical Accuracy
  - “Suppliance in Obligatory Context” (SOC)
  - Is a particular grammatical form supplied in the context that you would expect it?

$$\frac{\text{number of correct suppliance in obligatory contexts} \times 2 + \text{number of misformations in OCs}}{\text{total OCs} \times 2}$$

She begged for forgiveness, **begs** also for mercy.



- Grammatical Accuracy
  - “Target-like use”

number of correct suppliance in obligatory contexts

number of obligatory contexts + number of suppliance in non-OCs

She begged for forgiveness, **begs** also for mercy.



# Qualitative or Quantitative?



# Working with data

- Work in pairs to analyse 7 – 8 scripts from the Examination for the Certificate of Proficiency in English (ECPE)
- Mark the multi-word sequences that you think are formulaic (cf Wray's definition)
- Group the scripts according to their use of formulaic sequences.
- What patterns do you see?



# A practical application

- Title:

“Documenting features of written language production typical at different IELTS band score levels”
- Aims:
  - Identify the defining characteristics of written language performance at each band score
  - Explore how these features of written language change from one band score to another
  - Explain the effects of L1 and writing task type on the different features of written language production.



- Characteristics of writing at different levels
- Features explored:
  - Cohesive devices
  - Vocabulary richness
  - Syntactic complexity
  - Grammatical accuracy
- Why?
  - To cover a range of key areas of language performance
  - Features represent key aspects of most language assessment scales
  - Previous research (e.g. Wolfe-Quintero et al., 1998) suggested these are productive developmental measures.



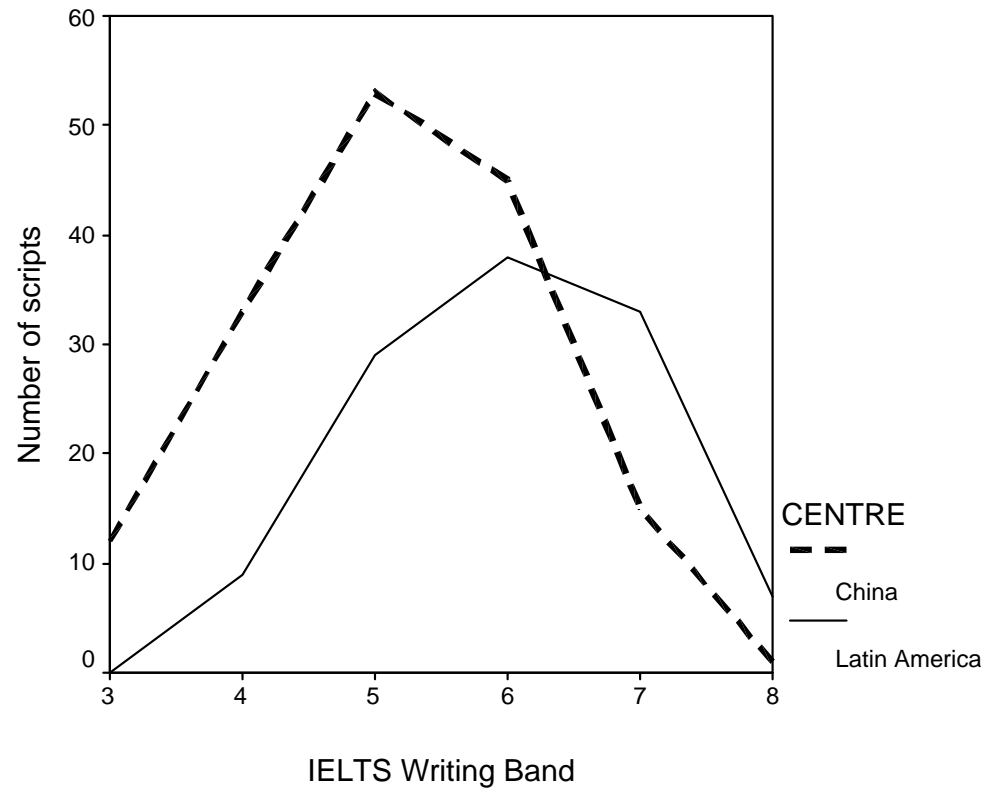
- Characteristics of writing at different levels
  - Health warnings (intervening factors):
    - L1 – SLA research (see Odlin, 2003) has shown that L1 has a clear and specific effect on L2 development.
    - Task – some evidence from earlier research (e.g. Mayor et al., 2002) that this would have an effect on performance.



- Characteristics of writing at different levels
  - All data provided by Cambridge ESOL
  - Sample
    - 275 test takers
    - L1 Chinese (159) + L1 Spanish (116)
    - Male (112) + Female (106) – evenly distributed between the two L1 groups
    - IELTS band scores 3 – 8
    - Tasks 1 and 2



- Characteristics of writing at different levels
- Distribution of scripts



- Characteristics of writing at different levels
- Concerns at the outset
  - The uneven distribution of scripts at each band score level.
  - The uneven distribution of performances for each L1 – that test-takers in China appear to take the IELTS test when they are at lower levels of L2 proficiency than test-takers in Latin America.
  - Lack of double-marking for the scripts – need to be cautious about claims that scripts have been accurately placed at particular band score levels.



- Characteristics of writing at different levels
  - Preparing the texts for analysis
    - Transcribed verbatim.
    - Each performance saved separately.
      - 2 tasks from each test-taker
    - Graphical errors that could interfere with the analysis were corrected.
    - Both the verbatim and the ‘corrected’ scripts were saved separately.



- Characteristics of writing at different levels

- Graphical aspects corrected:

- US spellings were changed to UK spelling, for consistency.
    - large numbers were written without the use of commas or spaces (e.g. 1000000).
    - repeated words were removed (e.g. the the).
    - obvious spelling errors (e.g. wrong vowel used).
    - words which were written as two separate words, but which should have been one, were joined (e.g. some body → somebody).
    - incorrectly formed past tenses (e.g. maked → made).
    - non-words were corrected to the nearest possible word, based on context (e.g. aniqilate → annihilate).
    - anagrams where the meaning was clear (e.g. from / form).
    - homonyms (e.g. there / their).



- Characteristics of writing at different levels
  - Aspects left uncorrected:
    - a word was used that was a real word but not the correct one (e.g. ‘insensible’ was left where ‘nonsensical’ would have been more appropriate).
    - plural / singular agreement was not corrected, as long as the forms were acceptable words (e.g. not ‘womens’).
    - grammatical errors.

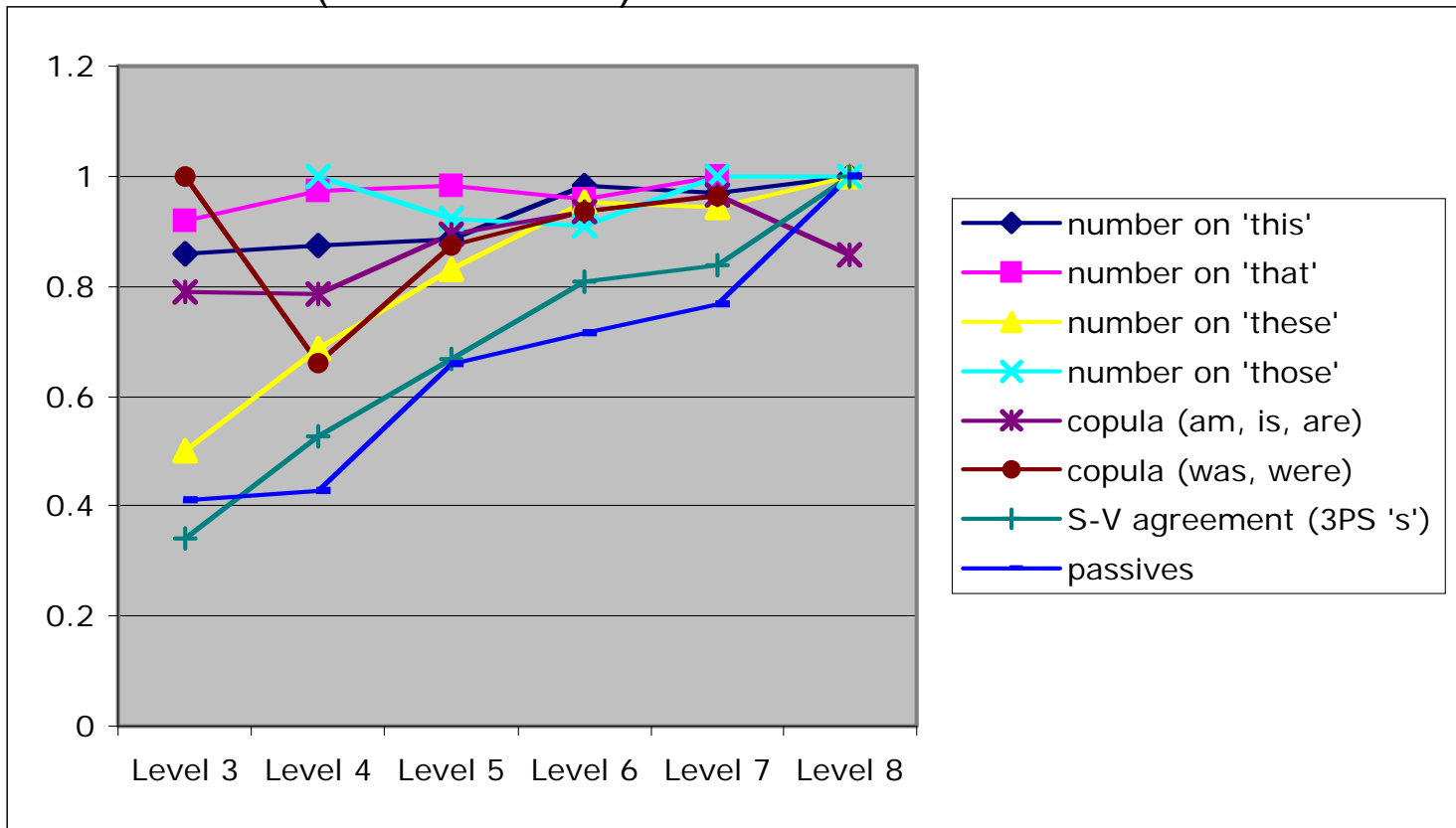


- Characteristics of writing at different levels
  - Some findings:
    - RQ1: Defining characteristics of performance at specific levels
      - ‘Early’ morphemes are accurate from early on, and ‘late’ morphemes get more accurate as L2 proficiency increases (e.g. TLU scores)



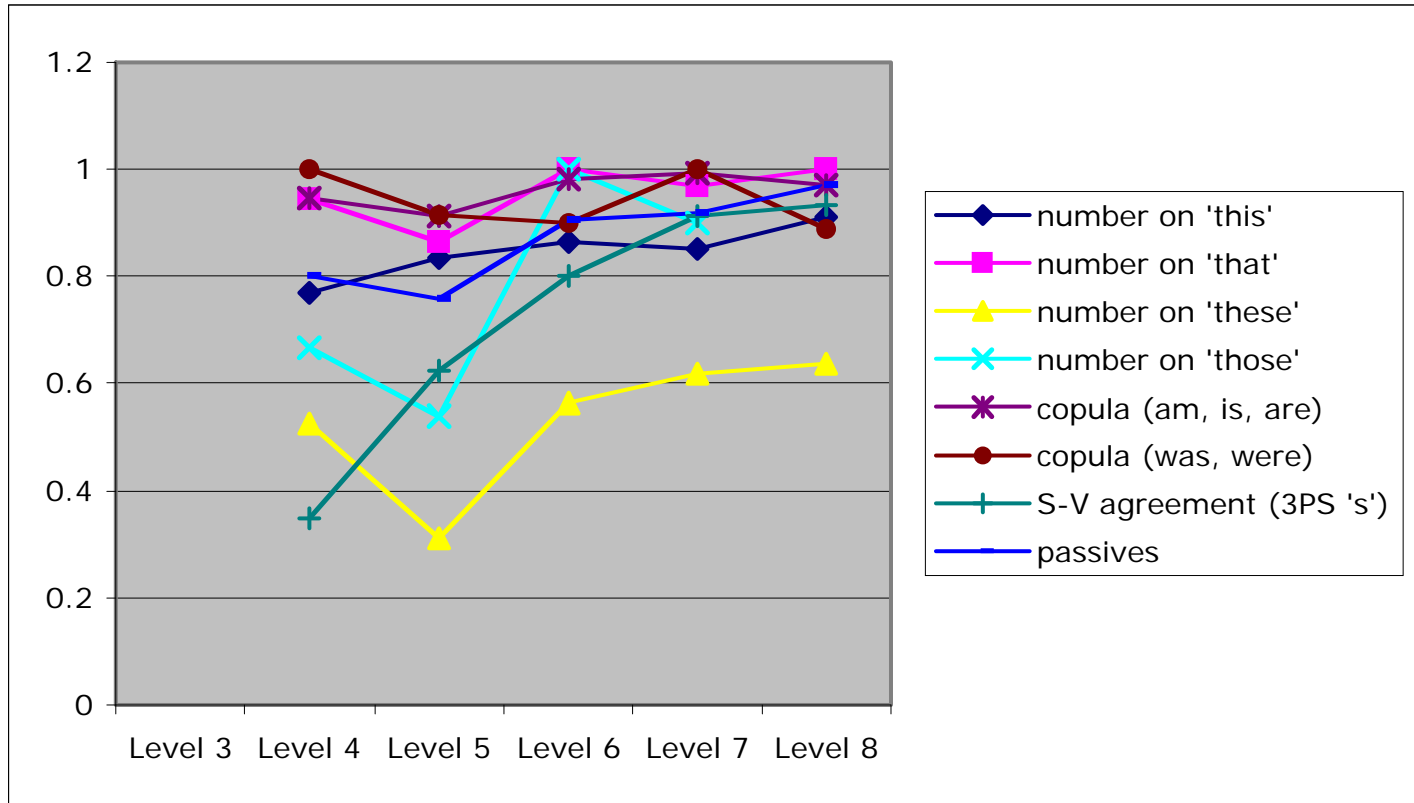
# •Characteristics of writing at different levels

•L1 Chinese (TLU measure)



# •Characteristics of writing at different levels

•L1 Spanish (TLU measure)



- Characteristics of writing at different levels
  - Some findings:
    - RQ2: How do features change from one level to the next?
      - There appears to be a link between the band level and the number of types and tokens produced.
      - High standard deviations within bands suggests variation in the number of types and tokens produced by test-takers within bands.



	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)



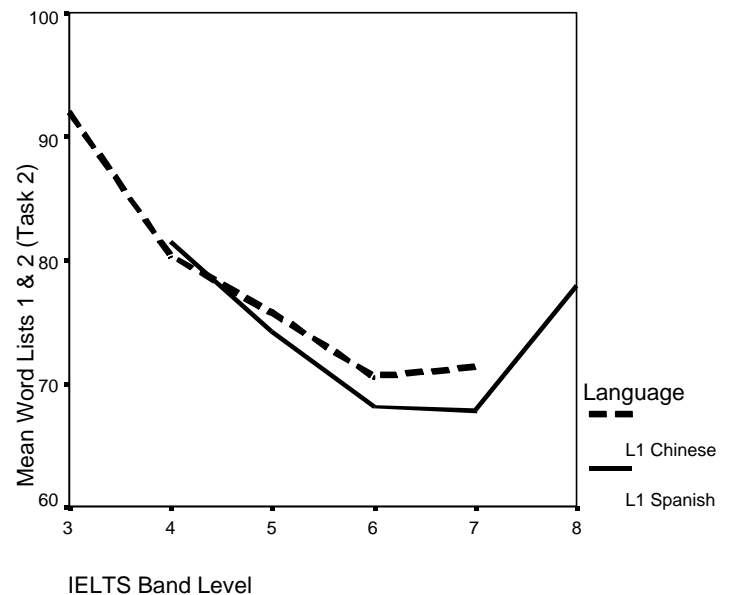
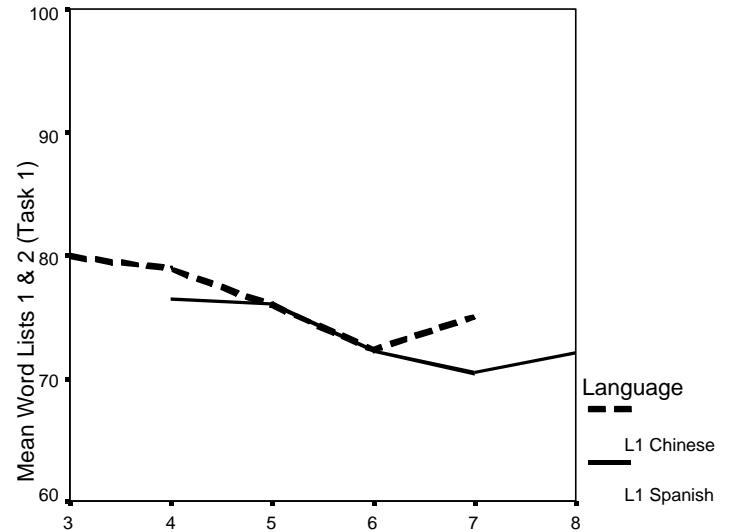
- Characteristics of writing at different levels
  - Some findings:
    - RQ2: How do features change from one level to the next?
      - 2-way ANOVA used to test the main effects and potential interactions of band level and L1
        - » L1 does influence the number of tokens and types produced (the L1 Spanish group tended to produce more tokens and types than the L1 Chinese group)
        - » The number of tokens and types produced does contribute to a test-taker's IELTS band level
        - » BUT – L1 does not interact significantly with Band level in terms of the number of tokens and types produced
        - » This confirms that there is no consistent pattern in the relationship between band score awarded and L1 in terms of the number of tokens and types produced.



- Characteristics of writing at different levels
  - Some findings:
    - RQ3a: L1 effects
      - Different L1 groups peak at different times e.g. measure of lexical sophistication (mean % of types from word lists 1 & 2 for both L1 groups in Task 1)
      - As test-taker's band scores increase they are less likely to use high-frequency words
      - When test-takers reach a critical IELTS band score (around IELTS band 6.0 for L1 Chinese speakers and around IELTS band 7.0 for L1 Spanish speakers) the pattern reverses. This could indicate a point at which a criterion other than vocabulary becomes more salient for assigning a level.

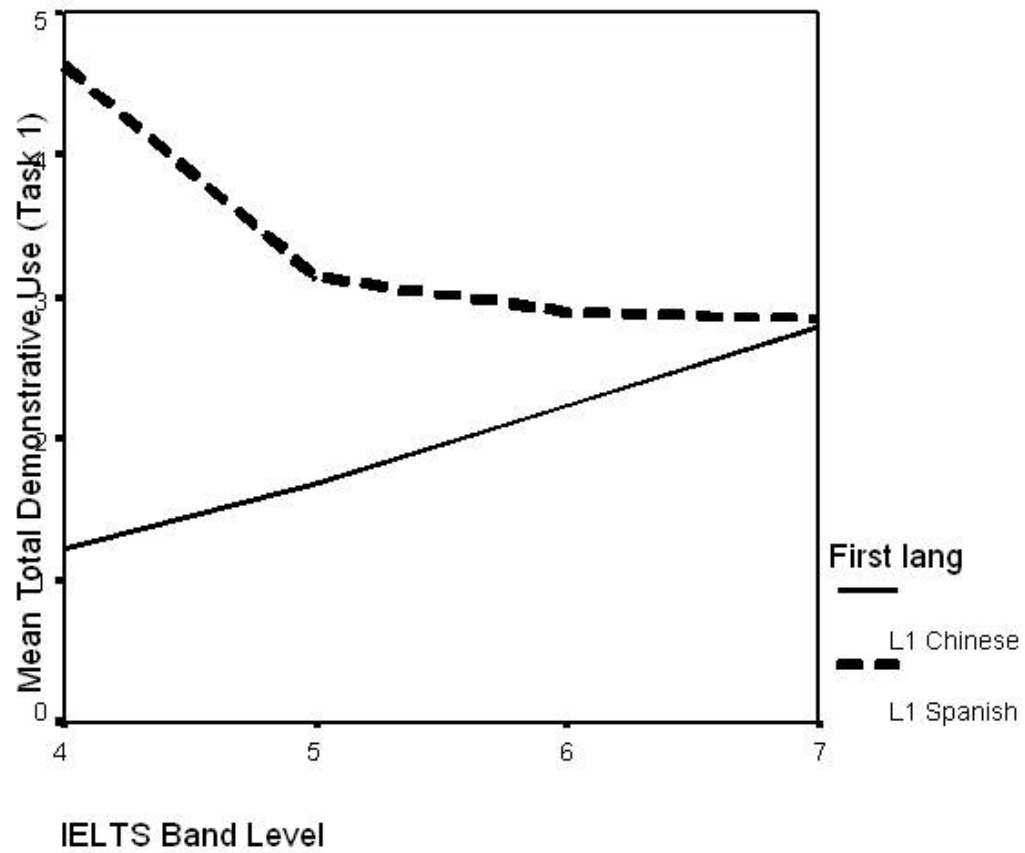


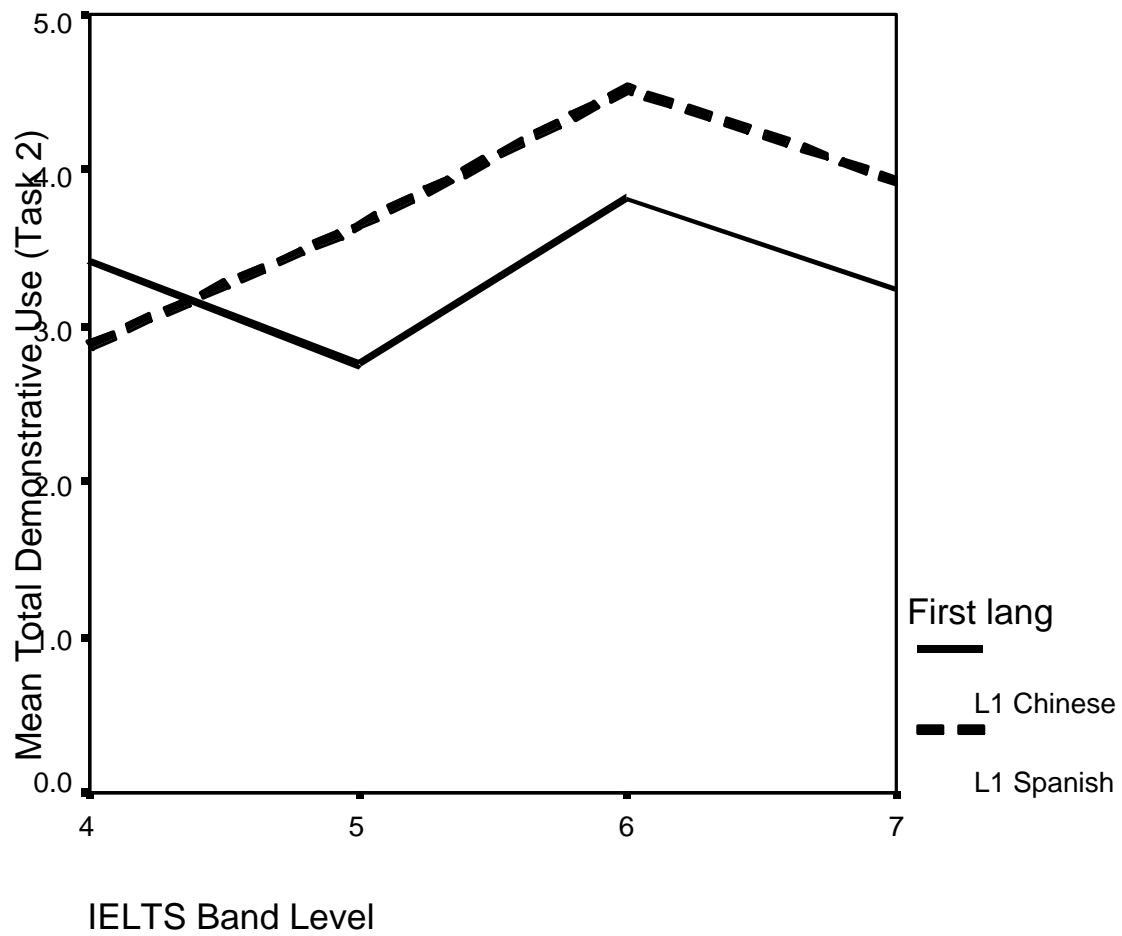
- Comparing the L1 groups:
  - The use of high frequency words declines more gradually in Task 1 than in Task 2.
  - Test-takers at lower IELTS band levels use considerably more high-frequency words for Task 2 than for Task 1.
  - For each L1 group there is a point at which the use of high-frequency words begins to climb again. This pattern occurs in both tasks.



- Characteristics of writing at different levels
  - Some findings:
    - RQ3b: Task effects
      - Different tasks can produce quite different profiles (e.g. mean frequency of use of demonstratives)







- Characteristics of writing at different levels
  - Aims:
    - Identify the defining characteristics of written language performance at each band score
    - Explore how these features of written language change from one band score to another
    - Explain the effects of L1 and writing task type on the different features of written language production.
    - Investigate how features of written language production interact at different performance levels



- How features of writing interact
  - Dataset reduced to bands 4-7 (255 test takers)
  - PCA/EFA to reduce measures in each language area so as to end up with measures that tap different aspects of proficiency in each of the 4 language areas
  - Ordinal regression analysis to uncover any existing relations between measures in the different language areas as proficiency develops



- How features of writing interact
  - Measures retained after PCA/EFA
- Cohesive devices:
  - Ratio of all demonstratives/tokens
- Vocabulary richness:
  - Lexical variation/density
  - Weighed lexical density
  - Lexical sophistication
- Syntactic complexity:
  - Dependent clauses per clause
- Grammatical accuracy:
  - Target-like use



- How features of writing interact
  - Why use ordinal regression?
- To find out how these measures interact in predicting band level
- Because we cannot assume linearity or interval measurement



- How features of writing interact
  - Procedure for obtaining ordinal regression
- Ordinal outcome variable: band level
- Predictor variables: 6 measures retained after PCA/EFA
- Step-wise backward regression
- Procedure done separately for each task



# Ordinal regression results

## Task 1

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[band = 4]	2.413	.847	8.122	1	.004	.754	4.073
	[band = 5]	4.335	.878	24.382	1	.000	2.614	6.055
	[band = 6]	6.116	.911	45.031	1	.000	4.329	7.902
Location	Lexical density	2.246	.948	5.611	1	.018	.388	4.105
	Syntactic complexity	2.914	.754	14.948	1	.000	1.437	4.392
	Grammatical accuracy	2.947	.684	18.581	1	.000	1.607	4.287
	[L1=1 ]	-.915	.257	12.729	1	.000	-1.418	-.413
	[L1=2 ]	0(a)	.	.	0	.	.	.

Link function: Logit.

a This parameter is set to zero because it is redundant.



# Ordinal regression results

## Task 2

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[band = 4]	-14.370	3.190	20.294	1	.000	-20.622	-8.118
	[band = 5]	-12.091	3.160	14.636	1	.000	-18.285	-5.897
	[band = 6]	-9.954	3.125	10.149	1	.001	-16.078	-3.830
Location	Lexical diversity	-.051	.021	5.697	1	.017	-.093	-.009
	Lexical sophistication	-.189	.029	42.677	1	.000	-.246	-.132
	Syntactic complexity	3.709	1.247	8.840	1	.003	1.264	6.154
	Grammatical accuracy	5.996	1.024	34.280	1	.000	3.989	8.003
	[L1=1 ]	.014	.290	.002	1	.960	-.554	.583
	[L1=2 ]	0(a)	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.



- How features of writing interact
  - Ordinal regression results
- Best predictors of band level regardless of task:
  - Syntactic complexity
  - Grammatical accuracy
- Best predictors of band level for Task 1:
  - Lexical density
  - Syntactic complexity
  - Grammatical accuracy
- Best predictors of band level for Task 2:
  - Lexical diversity
  - Lexical sophistication
  - Syntactic complexity
  - Grammatical accuracy



- How features of writing interact
  - Conclusions: Task 1
- The model explains bands 4 – 6 but not band 7
  - Task does not give higher level test-takers latitude to show what they can do while simultaneously supporting lower level test-takers.
  - For this task the measures that we have investigated are not relevant for higher level test-takers.
- The model accounts for a relatively modest proportion of the variability in the band levels.



- How features of writing interact
  - Conclusions: Task 2
- The model explains all 4 bands
- The model accounts for a higher proportion of the variability in the band levels (compared to Task 1).
  - Because more of the measures contribute to the model



- How features of writing interact
  - Conclusions:
- The cohesion measure used here does not seem to contribute to the model. However this is likely to be an artefact of the measures we conducted.
- Vocabulary measures contribute to both Task 1 and Task 2 BUT the measures are different.
  - Lexical density might be more relevant in Task 1 because of the lexical support provided by the input material.



- How features of writing interact
  - Final thoughts: plenty to do:
- More robust measures of coherence/discourse features.
- Taking a more grounded approach and looking for clusters in the data – looking for ‘natural bands’.
- Explore the interactions band by band i.e. what defines a Band 4?

