Adult test takers' lexical development:

A study of repeaters of a standardized writing proficiency test

You-Min Lin and Michelle Y. Chen
Paragon Testing Enterprises,
Vancouver, Canada

2020-10-31

CELPIP® CAEL® | by Paragon

Background

- Lexical sophistication is an important indicator of writing proficiency (Engber, 1995, Guo et al., 2013; Kyle & Crossley, 2016, Laufer & Nation, 1995)
- Repeat test-taker performance from standardized language tests offers valuable empirical data to investigate lexical development
- Lin & Chen (2020) observed changes in some lexical features within 6 months of re-testing, as opposed to features in syntactic sophistication and cohesion, where change was not evident.
- This study further investigates lexical development of repeaters between attempts by incorporating additional lexical measures

Research Questions

Our goal is to explore how repeaters' lexical development is affected by the duration between tests and initial writing proficiency:

- Did the test takers' lexical features change over time?
- If so, which lexical features changed over time?
- Did the changes vary after a relatively short (30-40 days) vs. a longer period (90-180 days)?
- Did the changes vary according to test takers' initial proficiency?

Data Source

CELPIP-General (The Canadian English Language Proficiency Index Program), Writing Assessment

- Computer-delivered standardized test measuring general English language proficiency
- Two independent writing tasks:
  - an e-mail to a service provider
  - a response to a survey question
- Rated on four dimensions: content/coherence, vocabulary, readability, and task fulfillment

Test-Taker Sample

Repeat Test Taker Sample

- 562 adult test takers (21 to 66 years old); 60% male and 40% female
- Took CELPIP-General Writing at least 3 times
- Second attempt at 30-40 days to initial testing, third attempt at 60-180 days to initial testing

Corpus

- 3,376 writing samples (562 test takers x 2 writing tasks x 3 attempts)
- Prompts are not controlled for (Attempt 1 = 299, Attempt 2 = 307, Attempt 3 = 333)
- Total 618,078 words; average length 183.08 (SD=32.58) words; 3304 (98%) are longer than 100 words, and 2892 (86%) fall between 150 to 250 words

Study Design

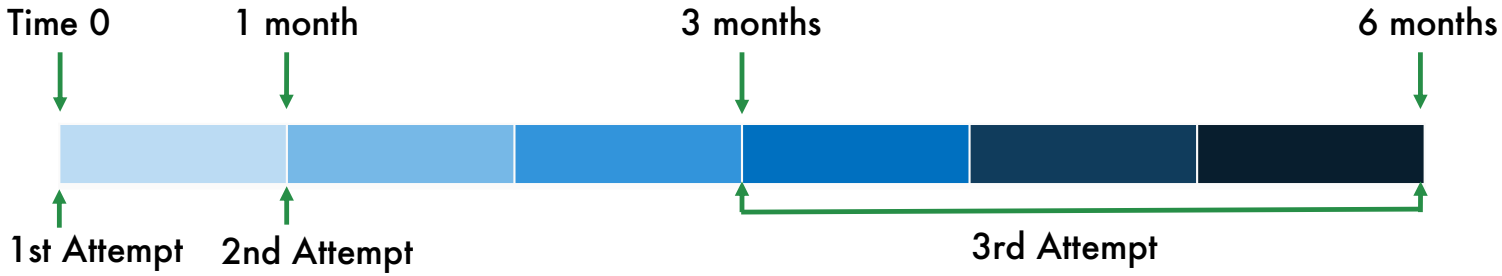
Adopted a 3 (Proficiency) x 3 (Attempt) mixed design.

- The 3 proficiency groups

Proficiency Group	Low		Medium			High	
CELPIP Level	3	4	5	6	7	8	9

  - Only test takers who scored above level 3 are included.

- The 3 attempts



Analysis

- Lexical features were analyzed using natural language processing (NLP) tools
- Mixed ANOVAs were conducted to examine the effect of the following parameters on linguistic features:
  - duration (i.e., attempt)
  - proficiency group
  - their interaction
  - *significant when p <.005*
- Post hoc tests were conducted when one or more of the main effects or the interaction are statistically significant:
  - *significant when p <.05*

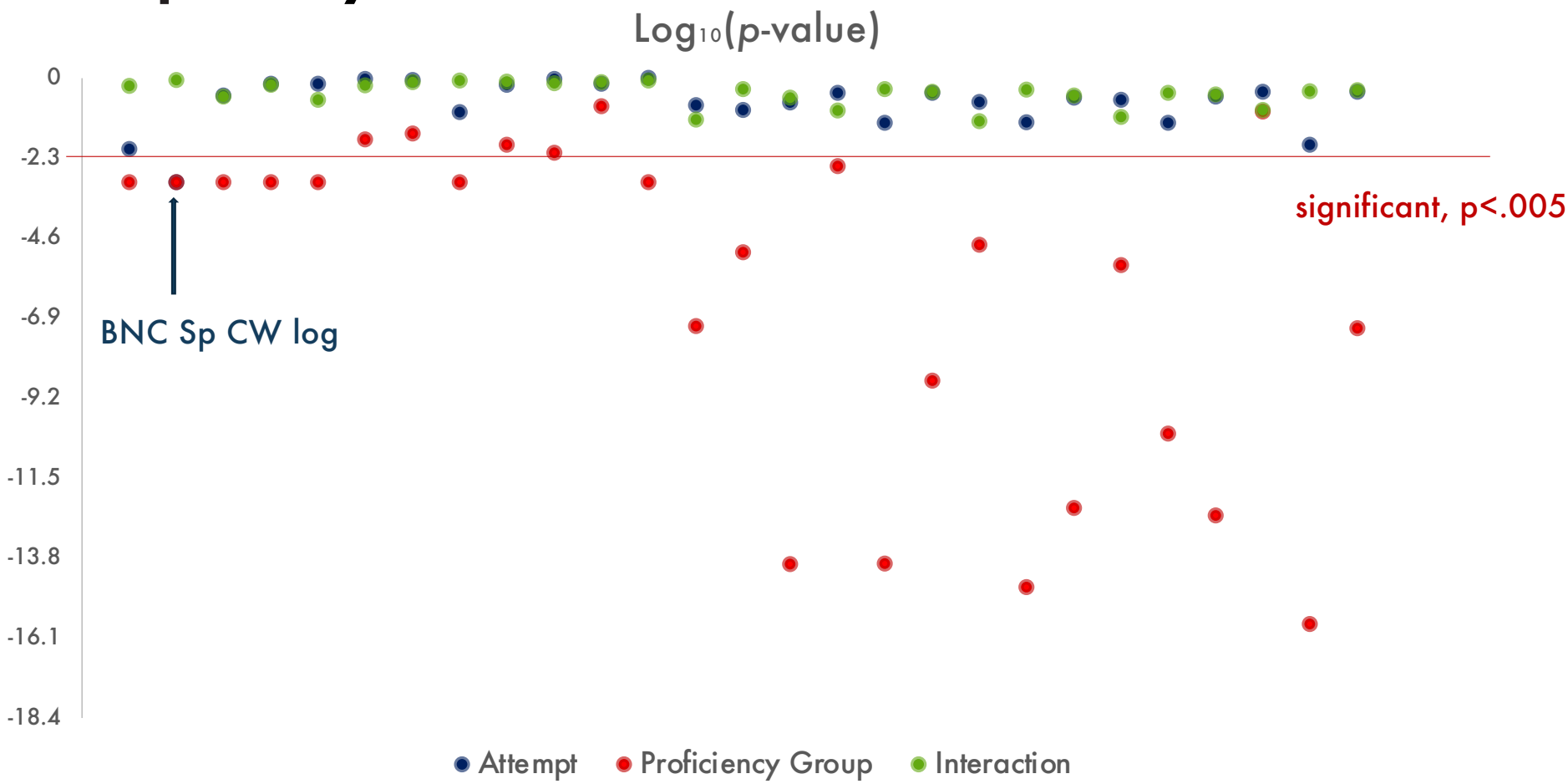
Lexical Sophistication

- Subsumes several constructs, often operationalized as a set of features:
  - lexical frequency – frequency in the reference corpus
  - lexical range – number of texts in the reference corpus a word/expression occurs
  - lexical diversity – type-token-ratio (TTR) and variants
  - semantic relationships – number of polysemous senses and hypernyms
  - psycholinguistic properties – properties of words based on human ratings and psycholinguistic experiments
- These features were measured by NLP tools (TAALES, Kyle & Crossley, 2015; Python lexical diversity package, Stanza, NLTK), which allowed for comparisons among large collections of textual data.

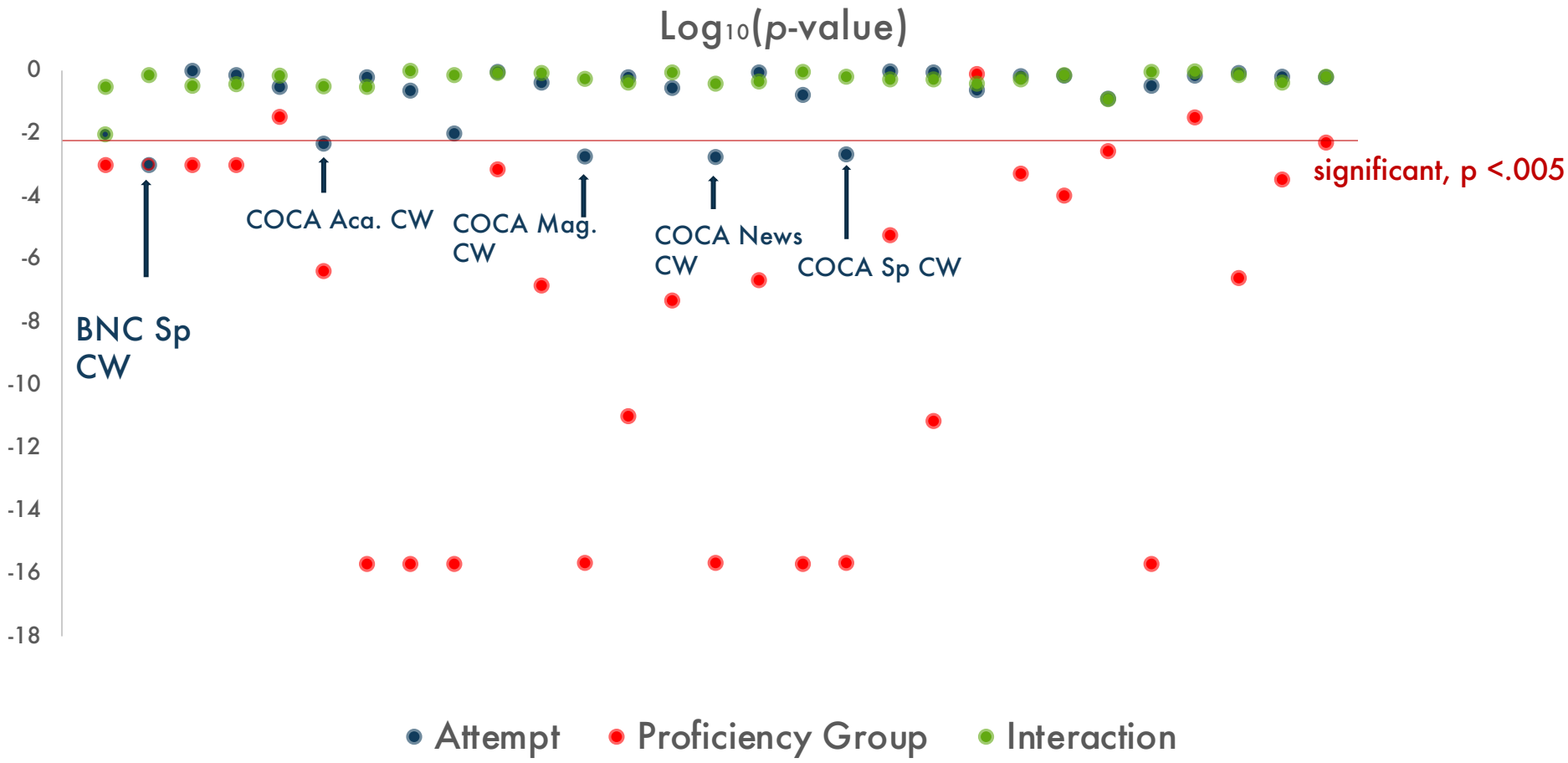
Lexical Indices

Type	Indices	
Frequency (27)	• BNC Spoken and Written Corpus • COCA Spoken and Written Corpus (academic, magazine, fiction, news)	All | content | function words Bigram & Trigram *(TAALES, Kyle&Crossley, 2015)*
Range (29)	• BNC Spoken and Written Corpus • COCA Spoken and Written Corpus (academic, magazine, fiction, news)	All | content | function words Bigram & Trigram *(TAALES, Kyle&Crossley, 2015)*
Diversity (5)	TTR, MSTTR, HDD, MATTR, MTLD	All words *(Python NLTK and lexical Diversity Package 0.1.0)*
Semantic Relationships (8)	Polysemy score (Wordnet synsets) Hypernymy score (Wordnet hypernym paths)	All words, verbs, nouns, adjectives, adverbs *(Python Stanza and NLTK Package for POS tagging and extraction of WordNet synset and hypernymy information)*
Psychological Properties (7)	MRC database familiarity, concreteness, imageability, meaningfulness (Coltheart, 1981).	Content | function words *(TAALES, Kyle&Crossley, 2015)*

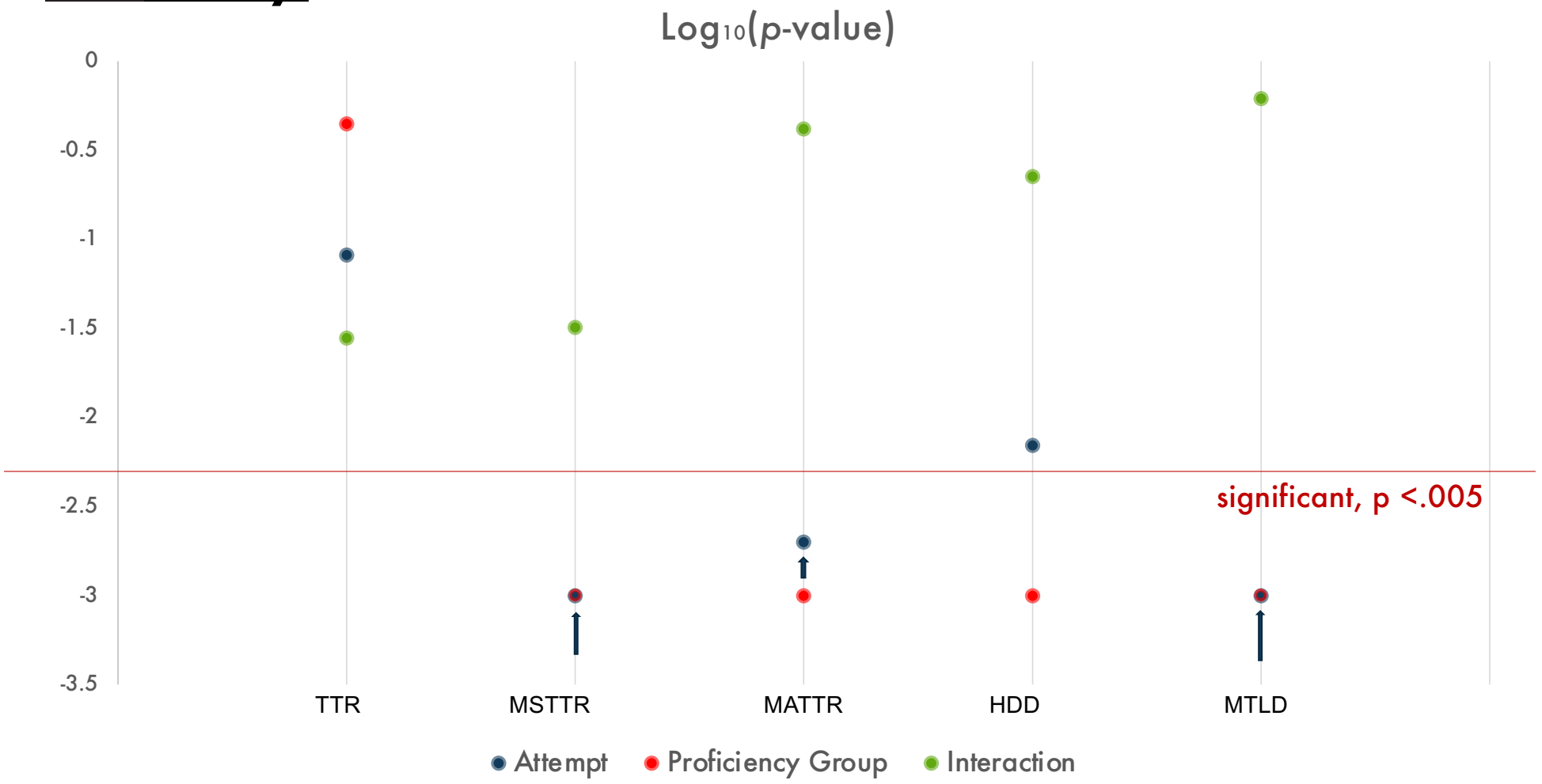
Frequency


Log10(p-value)
0
-2.3
-4.6
-6.9
-9.2
-11.5
-13.8
-16.1
-18.4
significant, p<.005
BNC Sp CW log
Attempt
Proficiency Group
Interaction

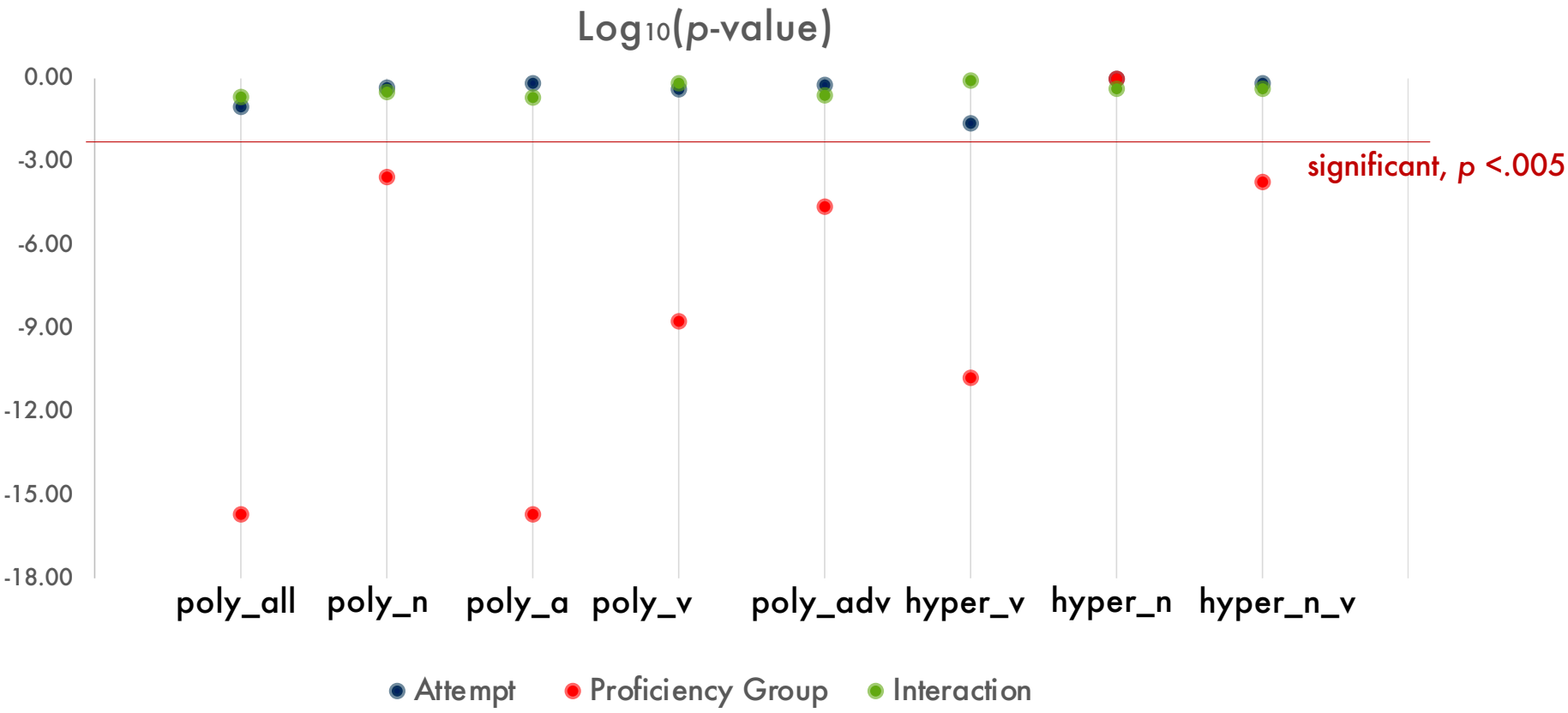
Range


Log10(p-value)
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
significant, p <.005
BNC Sp CW
COCA Aca. CW
COCA Mag. CW
COCA News CW
COCA Sp CW
Attempt
Proficiency Group
Interaction

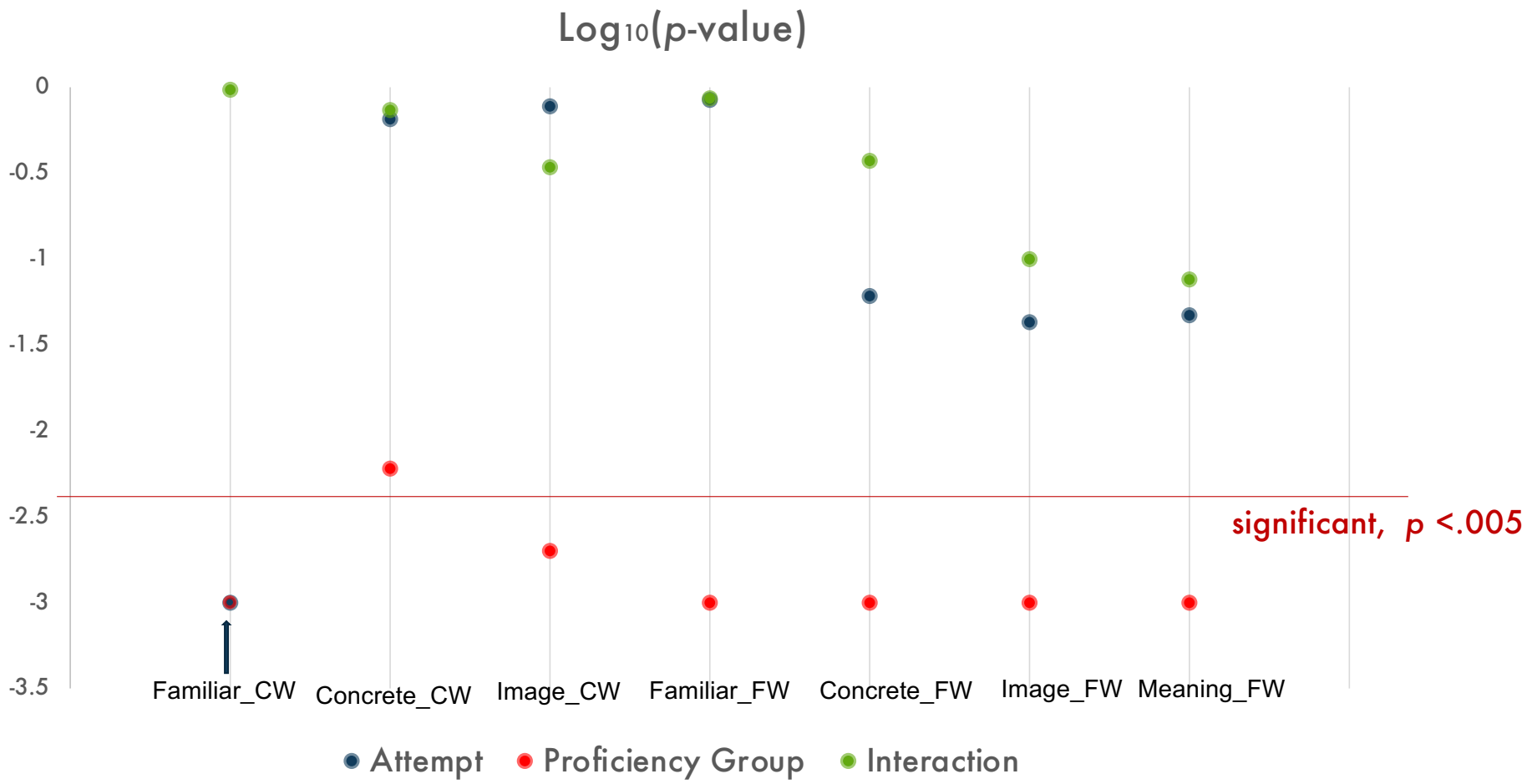
Diversity


Log10(p-value)
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
significant, p <.005
TTR
MSTTR
MATTR
HDD
MTLD
Attempt
Proficiency Group
Interaction

Semantic Relationships



Psychological Properties



Development Across Different Durations (All Proficiency Groups)

			1 month	3-6 months
Frequency	BNC CW log	Spoken	↓	↓
Range	BNC CW	Spoken	↓	↓
	COCA CW	Written (Academic)		↓
		Written (Magazine)		↓
		Written (News)		↓
		Spoken		↓
Diversity	MSTTR			↑
	MATTR			↑
	MTLD			↑
Psychological Properties	MRC CW Familiarity		↓	↓

Major Trends

- Limited lexical development within 6 months of re-testing
  - Relatively few indices show effect of attempt
- Significant effect of *proficiency groups* for many of the lexical indices analyzed.
  - 22 out of 27 frequency
  - 25 out of 29 range
  - 4 out of 5 diversity
  - 7 out of 8 semantic relationships
  - 6 out of 7 psychological properties
- This means that differences in lexical features are mainly associated with proficiency group, but there is relatively little development across attempts.
- No interaction: Observed across-attempt lexical development does not have different effects by proficiency group

Observed Change

The few changes observed:

- **Word Frequency (↓)**
  - Use of lower frequency, more sophisticated words
- **Word Range (↓)**
  - Use of domain & genre-specific vocabulary with a narrower range of distribution
- **Lexical Diversity (↑)**
  - Less repetition and more variety in vocabulary
- **Psychological Property (↓)**
  - Use of less familiar words

Discussion

Let's revisit our research questions:

- Did the test takers' lexical features change over time?
  - Yes, but change is only observed for some lexical features, not many
- If so, which lexical features changed over time?
  - Frequency, Range, Diversity, Psychological Properties
- Did the changes vary after a relatively short (30-40 days) vs. a longer period (90-180 days)?
  - Yes, depending on the individual features and reference corpora, some feature changes began at 1 month, while for others, changes were only observed at 3-6 months
- Did the changes vary according to test takers' initial proficiency?
  - No, the observed across-attempt changes did not differ for different proficiency groups

Future Plan

- Investigate the effect of reference corpora on the results for the selected measures
  - Both frequency and range indices may be affected by the composition of texts in the corpus
- Consider coverage of words (% occurring in reference corpus)
  - Non-matched words may affect the accuracy of measures and interpretation of results
- Add additional lexical indices (e.g., academic word lists, dispersion indices, association strengths) to increase construct representation
- Incorporate demographic information and language background as variables