

An analytic investigation of paired compositions from a high-stakes writing test
Barbara Dobson, Fernando Fleurquin, Aaron Ohlrogge, Tian Song, and Sarah VanBonn
ELI-University of Michigan PairedCompositionStudy@umich.edu
EALTA, Athens, Greece, May 2008

Background: 11,304 intermediate EFL learners chose to retake the writing section of a high-stakes examination, generating pairs of compositions on two different topics (Topic 1 and Topic 2).

Research Questions:

- Is the discourse generated in response to Topic 1 similar to/different from the discourse generated in response to Topic 2? If so, how?
 - For candidates who scored the same on T1 and T2
 - For candidates who scored differently
- How do various aspects of discourse relate to composition grade?
- What implications do the findings have for test developers, test takers, and language teachers?

Methodology:

- Sample of 110 pairs of compositions
 - 60 pairs of “differents” (Time 1 score ≠ Time 2 score)
 - 50 pairs of “sames” (Time 1 score = Time 2 score)
- Select measures to describe lexical output, lexical variation, lexical sophistication, accuracy, and original language/content
- Determine values on each measure for each of the 220 compositions
- Use 2-way ANOVA to compare these measures across topic and across score
- Use multinomial logistic regression to investigate the effect of these measures on composition scores

Discourse Measures:

MEASURES	DEFINITION	SOURCES/ INTERPRETATION
Lexical output	Number of tokens; word count	Compleat Lexical Tutor Vocabprofile http://lxtutor.ca/
Lexical variation	Type/Token Ratio corrected for length of text	$TTR2 = \text{types} / \sqrt{2 \cdot \text{Tokens}}$
Lexical sophistication	Percent of total tokens that are words in the 1,000 (K1) most-frequently used English words	Compleat Lexical Tutor Vocabprofile [higher values represent less sophistication--a higher proportion of common vocabulary]
Accuracy	Percent of total T-units that are error-free T-units	Coded by researchers
Original language	Percent of total tokens that are tokens given in the prompt	Code written by S. Wulff using www.r-project.org [higher values represent a lower proportion of original language]

Preliminary Results:

Lexical Output

- Same group: No significant difference in mean length between T1 and T2.
- Different group: Mean length is significantly shorter at T1 than at T2.
- For each topic, longer compositions receive significantly higher marks.

Lexical Variation

- Same group:
 - Mean lexical variation is significantly different for T1 and T2 but at only .045.
 - Different topics seemed to affect lexical variation for test takers with the lowest and highest scores but not for test takers with scores in the middle.
- Different group: No significant difference in mean lexical variation for T1 and T2.
- For each topic, lexical variation is significantly different across grade level.

Lexical Sophistication

- Both groups for both topics: the great majority of the language (>85%) is from the 1,000 most common words in English (K1).
- Both groups: T2 generated responses with a significantly higher % of K1 words.

- The mean % of K1 words was not significantly associated with grade.

Accuracy

- Both groups: There was no significant difference in the mean % of total T-units that are error-free.
- Both topics: The proportion of total T-units that are error free is significantly associated with grade.

Original Language

- Both groups: Responses to T1 contain a greater mean proportion of language from the prompt than do responses to T2.
- Prompts: For T2, the proportion of language from the prompt is significantly associated with grade. T1 does not show this pattern.

Multinomial Logistic Regression

- For both topics and both groups, 2 factors were found to significantly contribute to grade:
 - length (word count)
 - accuracy (% of correct T units)
- Controlling for other variables, lexical variation (TTR2) and lexical sophistication (%K1) did not significantly affect grade. Results for originality (% Tokens from Prompt) were mixed.

Implications for Test Developers:

A good prompt at this level is one that allows candidates to:

- Elaborate adequately on the topic
- Use general language to communicate their ideas
- Develop the topic without the need to excessively repeat the language of the prompt

Implications for Test Takers:

- Develop your ideas as fully as possible, trying to use more than just the language from the prompt and generic responses.
- Accuracy is important. Use language you can control.
- You do not need to use difficult or unusual words to pass.
- A very short composition is not likely to pass.

Implications for Teachers:

Help students . . .

- address and elaborate on the topic appropriately;
- improve accuracy;
- respond to prompts using their own words, as much as possible;
- write enough text in a timed situation to demonstrate control of language beyond that of the prompt.

Limitations:

- T1 were all letters/T2 included some essays. Genre difference was not controlled for.
- There were no A/A pairs of compositions.
- Measures used to analyze the discourse have limitations.
- The “different sample was selected to include all pairs with maximum score change; findings may not be generalizable to the entire population of retakers.
- The retakers were different than the total population who took this test (generally lower in ability on all sections of the test); findings may not be generalizable to entire population of test takers.
- All in the sample and in the test population have the same L1; findings may not be generalizable to all intermediate-level writers.

References:

- Cobb, T. (2007). *The Compleat Lexical Tutor* (v.6 11/07) Vocabprofile. <http://lextutor.ca/>
- Kuiken, F. & Vedder, I (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17 48-60.
- Lee, H-K & Anderson, C (2007). Validity and topic generality of a writing performance test. *Language Testing*, 24(3) 307-330.
- McCarthy, P.M. and Jarvis, S (2007). vocD: A theoretical and empirical evaluation. *Language Testing*, 24 (4) 459-488.
- O’Loughlin, K. & Wigglesworth, G. (2007) “Investigating task design in Academic Writing prompts” in L. Taylor & F. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing* (pp. 379-420).
- Read, J (2005). Applying lexical statistics to the IELTS speaking test. *Cambridge Research Notes*, 16, 12-16.
- “The R project for statistical computing.” <http://www.r-project.org>