Research Summary: Investigating lexical validity in the Pearson Test of English Academic

Kieran O'Loughlin
The University of Melbourne, Australia

December 2013

ABSTRACT

This report describes a study which examined the use of academic vocabulary in tasks requiring written production on a test of academic English designed to assess readiness for entry to English-medium higher education. The study examined the use of academic tokens (running words) and types (different words) in the stimuli and test taker responses to a wide range of items representing three different written production tasks: a summary based on a reading text, a summary based on a listening text and an essay. The holistic scores assigned to the essay responses were also used to examine the relationship between academic word use and writing proficiency. The results indicated a) highly variable use of academic tokens and types in the item stimuli and test taker responses between and within the three different tasks b) a significant relationship between the item stimuli and test taker responses in terms of academic tokens but not types and c) a strong correspondence between the holistic scores assigned to test taker essay responses and academic vocabulary use, in terms of tokens and, more particularly, types.

1. Introduction

An important quality of any academic proficiency test is what may be termed lexical validity: the extent to which the vocabulary occurring in, and elicited by, the test is representative of the vocabulary that test takers will encounter, and be expected to understand and produce, in real world academic contexts. To date, there seems to have been little or no research conducted to examine this question. The creation of the Academic Word List by Coxhead (2000) and the program RANGE (Heatley, Nation and Coxhead, 2002) now provide a valuable opportunity to investigate these issues.

The study of the effects of test tasks on test taker responses and on the test scores assigned to them is now an established area within language testing research (see Wigglesworth, 2008). Within this field of research the use of discourse analytic techniques to study test tasks and test taker responses is growing. This work builds on studies in Second Language Acquisition demonstrating that various task characteristics such as task structure, task difficulty and cognitive load, planning time and topic can affect learner task responses. The work of Foster and Skehan (Foster & Skehan, 1996; Skehan & Foster, 1997, 1999) has been particularly important in this regard. They showed that differences, such as whether a task is dialogic versus monologic, structured versus unstructured or simple versus complex in outcome may have a significant impact on measures of fluency, complexity and accuracy in the learners discourse. Recent studies in language testing have shown that fairly small variations in the

task can influence the output. For instance, using data from a test of speaking proficiency, Wigglesworth (1997) found that planning time resulted in measurable improvements in the complexity, fluency and accuracy of speech (although this was not reflected in the scores assigned by raters). There also appear to be format effects on test tasks in relation to test taker language output. O'Loughlin (2001) showed that comparable tasks used in direct and semi-direct formats of speaking tests showed significantly higher levels of lexical density in the responses on the semi-direct format. The influence of the interlocutor in task performance in direct tests or oral proficiency has also been explored. Brown (2003), for example, has demonstrated that the same test taker may produce qualitatively different performances on the same task when paired with different interlocutors.

While most of the studies in test taker production to date are concerned with oral assessment, Wigglesworth (1999) examined writing by comparing test taker responses to recount and report tasks. She found that in report tasks candidates used more complex but less accurate language, and in recount tasks their language was less complex but more accurate.

O'Loughlin & Wigglesworth (2007) examined how task difficulty was affected in an information transfer test task which formed part of the IELTS Academic Writing module. While the analyses of test scores did not reveal any significant effects, their discourse analyses of test taker responses (including task fulfilment, coherence, cohesion, vocabulary, sentence structure, and repetition of key words) across different proficiency levels revealed that tasks providing less information (and fewer key words) elicited more complex language. However, the pattern was less clear in relation to accuracy.

As O'Loughlin and Wigglesworth (2007: 381-382) suggest, the question of whether different task prompts affect the quantity and quality of test taker responses has generated conflicting findings in different studies about whether topic affects language output. However, such studies have generally focused on the ratings assigned to the responses and there has been little investigation of the actual writing itself. The problem here is that rater interacts with not only test taker's writing but also the task itself. The rater may consider one task prompt or stimulus to be more or less difficult than another (in terms of topic or complexity) and compensate for this perceived relative difficulty in their scoring. Thus, test scores may not accurately reflect the quality of the test taker response – hence the need for analyses of the responses as well. Conversely, there are many classroom-based studies which look at the discourse of learner essay writing but the essays have not been rated. This is a significant omission because the analyses of test taker responses can be more illuminating if they are grouped into the levels of performance assigned to them by raters.

Two important studies which have examined the relationship between the vocabulary used in test taker written responses and the scores assigned to them are Engber (1995) and Laufer and Nation (1995). Engber's (1995) study was based on 66 ESL compositions rated on a global scale from 1-6. Scores were assigned to each essay by eight raters and were then examined in relation to the essay's lexical richness through lexical variation (with and without errors of lexical choice and lexical form, percentage of lexical error, and lexical density). Significant correlations were found between the holistic scores and lexical variation, both with and without errors included. This suggests that lexical errors

of both choice and form may have a significant impact on raters' overall judgment of the quality of timed essays. While several relevant dimensions of lexical richness were examined to assess vocabulary use in this study, it failed to explore other important features such as lexical frequency and range.

Laufer and Nation (1995) established the Lexical Frequency Profile (LFP) which measures the proportion of high frequency general service and academic words in free written production. In their study, 65 students, who were divided into three proficiency levels (low intermediate, intermediate and upper intermediate), each wrote two compositions which were later analysed in terms of their range of vocabulary use, using the unit of word family (word stem plus all closely related affixed terms). The analyses were carried out using a program called VocabProfile (the precursor to the RANGE program used in the current study) enabling the results to be reported in relation to a) the first 1,000 and b) second 1,000 most frequent words in English, c) the University Word List (UWL) (Xue and Nation, 1984) and d) other words. The results indicated that the percentage of word families from the first and second 1,000 most frequent word lists was highest for the least proficient learners, while the percentage of word families used from UWL was highest for the most proficient learners. They also found that the LFP correlated strongly with an independent measure of vocabulary knowledge (the Vocabulary Levels Test). Laufer and Nation (1995: 316) suggest that the findings are "in accordance with the concept of language proficiency which assumes that richer vocabulary is characteristic of better language knowledge". They conclude that the LFP is, therefore, a valid and reliable measure of lexical use in writing and that it sheds light on the factors which affect judgment of quality in writing. One question about this study, however, is whether the word family was the appropriate unit of analysis since it does not reflect the words actually used in the student responses.

The studies by Engber (1995) and Laufer and Nation (1995) have pushed forward our understanding of vocabulary use in free written production. However, they are both small scale studies and focus exclusively on the range of vocabulary used by test takers. Frequency of vocabulary use is another important indicator of lexical richness since it takes into account how often words are used, as opposed to how many different words are used.

In terms of test taker language output, and vocabulary use in particular, there is a need to learn more about the relationship between task prompts and test taker output, examining the frequency and range of vocabulary use in response to both different task types, as well as different prompts used for the same task type. Further studies with larger sample sizes may strengthen our understanding of the relationship between range and frequency of test taker vocabulary use and the overall quality of written production, as reflected by the holistic scores assigned to them.

2. The Academic Word List (AWL)

The Academic Word List or AWL (Coxhead, 2000) was built from a corpus of 3.5 million *tokens* (the total number running words) and more than 70 thousand *types* (the total number of different words) of written academic text by examining the range and frequency of words outside the first 2,000 most frequently occurring words in English, as described by West (1953) in his General Service List (GSL). The corpus included a wide range of texts from the academic domain,

including 158 articles from academic journals, 51 edited academic journals from the World Wide Web, 43 complete university textbooks or course books, 42 texts from the Learned and Scientific section of the Wellington Corpus of Written English, 41 texts from the Learned and Scientific section of the Brown Corpus, 33 chapters from university textbooks, 31 texts from the Learned and Scientific section of the Lancaster-Oslo/Bergen (LOB) corpus, 13 books from the Academic Texts section of the MicroConcord academic corpus and two university psychology laboratory manuals (Coxhead, 2000: 219-220).

In creating the AWL, words were classified according to unit of "word family" (stem plus all closely related affixed terms) to which they belonged. Thus, the word family whose stem word is "concept" would include "conceptual" and "conception". Words were selected for the AWL based on three criteria: a) the word families had to fall outside the first 2,000 most frequently occurring words in English as represented by West's (1953) GSL; b) a member of a word family had to occur at least 10 times in each of the main sections of the corpus and in 15 or more of the 28 subject areas; c) members of a word family had to occur at least 100 times in the corpus (Coxhead 2000: 221).

The AWL includes 570 word families that constitute a specialised vocabulary with good coverage of academic texts, regardless of the subject area. It has been divided into 10 rank-ordered sub-lists according to decreasing word family frequency. With the exception of sub-list 10, each sub-list contains 60 items. More than 94% of the words in the list occur in 20 or more of the 28 subject areas covered in the Academic Corpus. These subject areas fall under four main disciplines: Arts, Commerce, Law and Science. West's (1953) GSL represented 76.1% of the tokens in the Academic Corpus, while the AWL accounted for 10% of tokens in the corpus.

Coxhead & Nation, (2001) argue that a general academic vocabulary is worth acquiring since it is common to a range of different texts, accounts for a substantial number of words in academic texts, is not as well known as technical vocabulary and is the kind of vocabulary students need to use in their academic studies.

While the AWL has attempted to identify a core academic vocabulary to be used across disciplines, recent research by Hyland and Tse (2007) has questioned its generalisations and underlying assumptions. In the investigation of their own developed corpus, they found that, although all 570 word families from the AWL appeared, their coverage across disciplines, social sciences, sciences and engineering was not evenly distributed. Students in the sciences had less coverage than other disciplines and only 36 of the word families were evenly distributed across all three disciplines. This then begs the question whether an academic vocabulary actually exists if it cannot capture the vocabulary required across disciplines. Hyland and Tse (2007:243) argue that language use in the different disciplines is more complex than a general academic vocabulary allows and that it may be more appropriate to develop a more specialised vocabulary approach according to each discipline. While this may be true, to date the AWL still represents the most systematic attempt to capture the kind of vocabulary that a test of general academic proficiency is likely to elicit from prospective students of higher education for whom English is an additional language.

3. The Pearson Test of English Academic

The Pearson Test of English (PTE) Academic is a new international English language test for students applying to enter universities, colleges and other higher education institutions, as well as professional and government bodies that require academic level language mastery. It provides measures of reading, writing, listening and speaking ability of test takers who are non-native speakers of English and who want to study at institutions where English is the medium of instruction. The test is endorsed by the Graduate Management Admission Council (GMAC), the organisation responsible for the GMAT (Graduate Management Admission Test) and will be launched in 2009.

The PTE Academic includes 20 different integrated tasks as item types. A large bank of items has been developed to represent each task. The test is computer-based and of three hours duration.

In the current study the following four research questions were investigated:

1. To what extent do the item stimuli for the different PTE tasks requiring written production include academic vocabulary?
2. To what extent do these tasks elicit academic vocabulary in test taker responses?
3. To what extent are the test taker responses to these tasks related to the stimuli for each item in terms of academic vocabulary?
4. To what degree are the frequency and range of academic vocabulary in test taker responses to the essay item type related to the average global score assigned by human raters to these responses?

4. Method

4.1 Data collection

The following three PTE tasks requiring a written response formed the focus of this study:

The data used in the study was taken from the first field test conducted in 2007 which was undertaken by more than 6,000 candidates from 21 countries. They completed 38 overlapping subsets of 21 different tasks in a computer-mediated environment. Test-takers completed a total of 95 items within 195 minutes.

The following three PTE tasks requiring a written response were the focus of this study:

RW-SUMM (Summary based on a reading passage)

In this task test takers are required to synthesize information in a reading text and write a one-sentence summary.

LW-SUMM (Summary based on a listening passage)

In this task test takers listen to a monologic audio recording, and then write a summary in 50-70 words of what the speaker has said. The test taker can use notes when listening to the recording, and use these notes as a guide to write the summary.

WW-ESSA (Essay)

In this task test takers write a persuasive essay in 200-300 words and support their position or opinions with details and examples.

The instructions for these three tasks, as well as sample item stimuli and test taker responses for each task, are included as Appendix 1. The sample items and responses were provided to the researcher by the PTE developers as examples of the three different tasks which had been trialled in the first field test, but which would not be used in future. They were therefore not chosen on the basis of how well they represented the percentages of AWL tokens or types in the task stimuli (or responses).

The stimuli and test taker responses gathered for all items representing these three tasks, as well as the (rounded down) average scores of two trained human raters assigned to test taker responses to WW-ESSA from the first field test of the PTE Academic, conducted in 2007 constituted the main data for the study. The item stimuli and the responses were then analysed in relation to the Academic Word List (AWL) developed by Coxhead (2000).

5. Data analysis

The item stimuli and responses were analysed using the program RANGE (Heatley, Nation and Coxhead, 2002). This program provides a) frequency counts of *tokens*, *types* and *word families* (word stems plus closely related affixed forms) and b) percentages of token and types in the target texts in relation to four word lists: 1) the first 1,000 most frequent words in English and 2) the second 1,000 most frequent words in English, 3) academic words and 4) other words not accounted for in the three previous lists. The first two word lists taken together constitute the GSL (West, 1953). The third word list is the AWL (Coxhead, 2000) containing 3107 types belonging to 570 word families.

Each RANGE output also provides lists of individual types from each of the four word lists, together with their frequency in the task stimuli and test taker responses.

The RANGE results relating to the numbers and percentages of AWL tokens and types in the task stimuli and test taker responses were the foci of this study. The token count provides a measure of how often AWL words were used overall (frequency), while types provide a measure of how many different AWL words were used overall (range). They therefore yield complementary information. The figures for word families are not reported as the RANGE output does not indicate which family members were actually used in the texts.

The data collected to address this question was drawn from the first global field test of the PTE conducted in 2007. Initially, the combined responses for all items representing each of the three test tasks were run through the RANGE program. Table 1 shows a) the total number of test taker responses, b) the total number of all response tokens (including AWL and other tokens) across all items representing the three tests tasks examined in this study and c) the average number of tokens per response for each of the three tasks.

Table 1. Item types, total written responses and total response tokens (Field Test 1, 2007).

TASK	Items	Total responses	Total response tokens	Mean tokens
RW- CONC	38	11, 284	355,611	31.52
LW- SUMM	38	11, 414	603, 241	52.85
WW-ESSA	19	5,622	967,134	172.03

6. Results

Each of the four research questions are now addressed in turn:

1. To what extent do the item stimuli for the PTE tasks requiring written production include academic vocabulary?

For each of the three tasks the stimuli for the various items used in the field test were examined using the RANGE program. There were 38 items used for each of the RW-SUMM and LW-SUMM tasks, and 19 items for the WW-ESSA task. Each stimulus was analysed separately.

Table 2 shows the number of items representing each task, the mean percentages of AWL tokens, standard deviations and range of AWL tokens used across the stimuli for the different items representing each task. These figures therefore provide a picture of the overall frequency of AWL words across the three tasks.

Table 2. Item stimuli: descriptive statistics for AWL tokens

TASK	No of items	Mean AWL tokens (%)	Standard deviatic	AWL token range (%)
RW-SUMM	38	7.24	3.89	0.70-16.36
LW-SUMM	38	5.93	2.58	1.34-12.24
WW-ESSA	19	5.86	4.06	0-10.71

These findings indicate that, on average, the stimuli for the items representing the RW-SUMM task included more academic tokens than the items for the other two tasks. The standard deviation results indicate greatest variability in the percentage of AWL tokens across the WW-ESSA item stimuli than across the items representing the other two tasks, although the range of percentages is widest for the RW-SUMM task. The RW-SUMM task also has the highest maximum percentage of AWL tokens across its various item stimuli.

Table 3 below shows the number of items for each task, the mean, standard deviations and ranges for the percentages of AWL types across the different item stimuli used for each task. These figures provide a picture of the overall range of AWL words used across the three tasks.

Table 3. Item stimuli: descriptive statistics for AWL types

TASK	No. of items	Mean AWL types (%)	Standard deviation	AWL type range (%)
RW-SUMM	38	9.38	4.60	1.35-22.09
LW -SUMM	38	8.29	3.29	1.80-15.22
WW-ESSA	19	6.54	4.62	0-13.04

Similar to the trend for AWL tokens in Table 2, the mean percentage of AWL types for RW-SUMM and the range of percentages are wider than for the other two tasks. However, as in the token analyses, the WW-ESSA shows the highest standard deviation. And like the token analyses, the RW-SUMM task has clearly the highest maximum percentage of AWL types across its various item stimuli.

2. To what extent do these tasks elicit academic vocabulary in the test taker responses?

In order to address Research Question 2, test taker responses to each of the items representing the three tasks were combined for the RANGE analyses. While this meant that information about the responses of individual test takers was lost, the main aim here was to compare the overall frequency and range of AWL tokens and types in the responses across a) the three tasks and b) the items representing each task and b) between the three tasks.

Table 4 below summarizes the findings from these analyses in terms of AWL tokens used in the responses.

Table 4. Test taker responses: descriptive statistics for AWL tokens

TASK	No. of items	Mean AWL tokens (%)	Standard deviation	AWL type range (%)
RW-CONC	38	7.589	3.904	2.04 -17.98
LW-SUMM	38	6.368	2.120	2.24 -13.92
WW-ESSA	19	5.366	1.113	4.05 -7.24

The results here suggest that the RW-SUMM task overall elicited the most AWL tokens. However, the standard deviation and range figures indicate that there was greater variation in the percentages of AWL tokens for each individual item than for the other two tasks. The maximum percentage of response tokens is clearly lowest for the WW-ESSA task.

Table 5 below summarizes the findings from these analyses in terms of AWL types used in the responses.

Table 5. Test taker responses: descriptive statistics for AWL types

TASK	No. of items	Mean AWL types (%)	Standard deviation	AWL type range (%)
RW-SUMM	38	12.24	2.45	7.7- 16.71
LW-SUMM	38	10.96	2.25	4.68- 14.86
WW-ESSA	19	12.92	1.24	11.07 -15.79

Here WW-ESSA elicits the highest average number of AWL types and less variability in the percentage of different AWL words across the individual items than the other two tasks, as indicated by both the standard deviation and range figures. It is notable that here the WW-ESSA also has the highest minimum percentage of AWL response types (11.07%) while the maximum percentages are very similar.

3. To what extent are the test taker responses to these tasks related to the stimuli for each item in terms of academic vocabulary?

Pearson ***r*** correlations were firstly calculated as a measure of the relationship between the percentage of AWL tokens in the task stimuli and the test taker responses for each task. As shown in Table 6 below, the co-efficients were all significant at the 0.01 level.

Table 6. Pearson *r* correlations for AWL tokens in the task stimuli and test taker responses.

TASK	No. of items	Pearson *r*
RW-SUMM	38	0.892 ($p < 0.01$)
LW-SUMM	38	0.679 ($p < 0.01$)
WW-ESSA	19	0.726 ($p < 0.01$)

The findings here suggest that the percentage of academic tokens in the written responses for all three tasks are quite strongly related to the percentage of academic tokens in the written and spoken stimuli provided as input, especially RW-SUMM.

Figures 1, 2 and 3 (see Appendix B) provide a graphical representation of the relationship for AWL tokens used in each of the three tasks. Each figure provides a comparison of AWL tokens for the stimulus and total test taker responses for each item representing the specified task expressed as percentages. The relationship between the item stimuli and the test taker responses is clearly strongest for RW-SUMM, confirming the correlation results in Table 6 above.

Correlations were also calculated for the AWL types in the task stimuli and the

test taker responses. The results are provided in Table 7 below:

Table 7. Pearson *r* correlations for AWL types in the task stimuli and test taker responses.

TASK	No. of items	Pearson *r*
RW-SUMM	38	0.714 ($p < 0.01$)
LW-SUMM	38	0.447 ($p < 0.01$)
WW-ESSA	19	0.127 (n.s.)

These results indicate that the relationship between AWL types in the task stimuli and test taker responses is weaker than for the AWL tokens, especially for the WW-ESSA task.

Figures 4, 5 and 6 (see Appendix B) provide a graphical representation of the relationship for AWL types used in each of the three tasks. Each figure provides a comparison of AWL types (expressed as percentages) for the stimulus and total test taker responses for each item representing the specified task. For all three tasks the percentage of AWL types is generally higher in the test taker responses than in the item stimuli. However, Figures 4 and 5 clearly indicate that the use of AWL types in the responses is still related to the use of AWL types in the stimuli for the RW-SUMM and LW-SUMM tasks but not for WW-ESSA. This may be because test takers understand that, in a test of academic English, the stimuli in the essay task are designed to prompt them to display the breadth of their academic vocabulary irrespective of how many AWL types appear in the prompt.

The limitation of the results to this point is that they only provide an overall picture of the relationship between test taker responses and the stimuli across all of the items representing the three tasks. They do not indicate, for example, the actual frequencies of AWL tokens and types in the stimuli and responses for each item. Nor do they show the relationship between the stimuli and responses on individual items in terms of tokens and types. In order to provide a closer view of the RANGE results, the results for the three sample items included in the Appendix are provided in Table 8 (tokens) and Table 9 (types) below. In the two tables the percentage of tokens or types for each sample item stimulus and set of responses is shown together with the actual frequencies on which the percentages are based. For example, the first cell of Table 8 indicates that 7.44 % or 9 of the 121 tokens in the RW-SUMM sample item stimulus were AWL tokens.

Table 8. Sample items: percentages of AWL tokens in the stimuli and responses.

SAMPLE TASK	AWL stimulus tokens	Responses (n)	AWL response tokens
RW-SUMM	7.44% (9/121)	218	8.26% (475/5,749)
LW-SUMM	12.25% (30/245)	215	13.92% (1741/12,504)
WW-ESSA	1.85% (1/54)	219	5.06% (2029/40,066)

The results for the RW-SUMM and LW-SUMM sample item are above average for

both stimuli and responses but those for the WW-ESSA sample item are below the mean in both cases (see Tables 2 and 4).

Table 9. Sample items: percentages of AWL types in the stimuli and responses.

SAMPLE TASK	AWL stimulus types	Responses (n)	AWL response types
RW-SUMM	10.26% (8/78)	218	13.98% (110/787)
LW-SUMM	14.77 % (22/149)	215	11.18% (189/1,691)
WW-ESSA	2.44% (1/41)	219	14.29% (583/4,080)

For AWL types the results for all three sample item are above average for both stimuli and responses (see Tables 3 and 5).

The above analyses still do not provide any information about which AWL words were used in the stimuli and responses. However, the output from each RANGE program analysis also routinely provides a list of individual AWL words as well as their frequencies in the individual item stimuli and the combined test taker responses. For this purpose, the results for the stimuli and combined test taker responses for these three sample items are reported.

RW-SUMM SAMPLE ITEM

As indicated in Table 9, there were eight AWL types occurring in the reading passage for this item. There were 110 AWL types used in the responses although the vast majority were used infrequently across the combined 218 responses to this item. As shown in the list below, there were only five AWL types occurring more than twenty times across the responses, with the key word "METHOD" by the far the most frequently occurring word.

AWL TYPE	RESPONSE FREQUENCY	STIMULUS FREQUENCY
METHOD	154	1
MINIMISE	37	1
PROCESS	37	1
REQUIRED	37	2
OUTPUT	27	1

There was only one AWL type used in the stimulus but not the responses ("CONSEQUENTLY"). Overall, 61.47% (292/475) of all AWL tokens used in the responses to this item were repetitions of words used in the item stimulus, and 52.74% (154/292) of these repetitions were "METHOD". The high rate repetition of AWL stimulus words (particularly "METHOD") in the responses underscores the nature of this task which is to briefly summarize the passage. It may be that repetition of key words (including AWL words) used in the stimulus is an important strategy for test takers in completing this task.

LW-SUMM SAMPLE ITEM

As indicated in Table 9, there were 22 AWL types used in the stimulus. There were 189 AWL types which occurred in the responses for this item, although (like the RW-SUMM item above) most were used infrequently across the combined 215 responses to this item. As shown in the list below, there were only 16 AWL types occurring more than twenty times across the total of 215 responses. 12 (75%) of these most frequently occurring AWL response types were also present in the stimulus listening text. In addition, four of the five most frequently occurring response types ("STABILITY", "ECONOMY", "ECONOMIC" and "DEPRESSION") were used more than once in the stimulus listening text.

Overall, 73.33% of all AWL tokens occurring in the responses to these items were repetitions of types appearing in the stimulus passage. These results indicate that test takers relied even more strongly on repetition of AWL types from the stimulus passage to successfully complete their short summaries than they did in the RW-SUMM sample item. They also repeated a broader range of AWL stimulus types, probably because there were more AWL types in the stimulus. However, the list above also indicates test takers used other AWL types not included in the stimulus passage such as "LECTURE, "INSTABILITY", "TOPIC" and "LECTURER".

Overall, only 4.14% (84/2029) of all AWL tokens occurring in the responses to this item were types (in this case a single type) appearing in the stimulus passage. The wider use of AWL types not appearing in this item stimulus by test takers, compared to the other two sample items, can be explained by the more strongly "productive" nature of the task here: the challenge to the test taker is less about manipulating language already presented to the test taker in the stimulus in order to build a short summary (as in the other two tasks), than to use the prompt as a springboard to write an essay demonstrating the ability to use a broader range of vocabulary (including AWL words) relevant to the topic.

AWL TYPE	RESPONSE FREQUENCY	STIMULUS FREQUENCY
STABILITY	247	5
ECONOMY	192	2
ECONOMIC	166	2
LECTURE	108	0
DEPRESSION	94	2
ECONOMICS	89	1
STABLE	72	1
INSTABILITY	66	0
ROLE	51	2
IMPACT	50	1
DISTRIBUTION	43	1
INCOME	39	1
TOPIC	30	0
RECOVER	25	1
CONCEPT	24	1
LECTURER	22	0

WW-ESSA SAMPLE ITEM

As shown in Table 9, there were 583 AWL types used in the responses, although (as for the other two sample items above) most were used sporadically across the 219 responses to this item. As shown in the list below, there were eleven AWL types which occurred more than twenty times across the responses. The only AWL word in the stimulus ("ELEMENT") was the most frequently used AWL word in the responses.

AWL TYPE	RESPONSE FREQUENCY	STIMULUS FREQUENCY
ELEMENT	84	1
ECONOMY	51	0
TECHNOLOGY	50	0
JOB	46	0
JOBS	33	0
INDIVIDUAL	32	0
RESOURCES	32	0
FACTOR	30	0
ROLE	30	0
ASPECTS	26	0
CONTRIBUTE	23	0

The results for these sample items should NOT be taken as representative of trends in the three different tasks. It is worth re-iterating that they were offered to the researcher simply as illustrations of the three different tasks by the PTE developers.

4. To what degree are the frequency and range of academic vocabulary in test taker responses to the essay task related to the average global score assigned by human raters to these responses?

In order to address this final research question, responses to the WW-ESSA task were grouped according to the average global score (adjusted downwards) derived from two independent human ratings between 0-4 (whole numbers only). These ratings were based on the scale for Overall Written Production from the CEF (Council of Europe, 2001, 61-62) such that 4= C2, 3 = C1, 2=B2, 1 = B1, 0 = <B1). This average score was available for a total of 4,956 of the 5,622 responses collected from the first field test.

Table 10 shows the percentage of AWL tokens and types for each score. In this instance, the figures for the percentage of AWL types have also been included since they provide useful additional information about the range of AWL individual words used by test takers at each score level.

Table 10. Percentages of AWL tokens and types for the Average Scores assigned to WW-ESSA responses.

Average score	No. of responses	% AWL tokens	% AWL types
4 (C2)	37	7.93	17.08
3 (C1)	285	6.42	15.21
2 (B2)	1654	5.72	10.22
1 (B1)	2319	4.76	8.59
0	642	3.76	9.32

Note that the percentages of AWL tokens and types systematically decrease as the scores descend. It is particularly striking that the percentage of AWL types decreases sharply from score 3 (C1) to 2 (B2). More generally, these results suggest that the frequency and, more particularly, breadth of academic vocabulary use appear therefore to be important markers of quality in the essay responses.

7. Discussion and Conclusion

The results indicated that the percentages of academic tokens and types in the stimuli of the items representing the three test tasks generally varied considerably. This suggests that the percentages of AWL tokens and types need to be monitored in the development process, so that the items representing each task present a similar level of challenge in terms of academic vocabulary to test takers.

In terms of the test taker responses, the obvious question is how much academic vocabulary should be evident in test taker responses. Nation (2008, personal communication) suggests that a figure of four percent or more of AWL tokens in responses to a given item indicates that a test item elicits an adequate level of academic vocabulary for a test of academic proficiency. From this perspective, as shown in Table 4 above, the mean figures for the use of AWL tokens in test taker responses on each of the three tasks provide adequate evidence for the lexical validity of the PTE Academic. However, the AWL token range figures indicate that responses to the test items representing each task did not always reach the minimum threshold of four percent.

Furthermore, the findings shown in Tables 6 and 7 indicate that there was a fairly strong positive correlation between the percentages of AWL tokens (but not types) in the stimuli and responses. This result suggests that the token relationship between stimuli and responses could be a fruitful focus for future test development work since the percentages of AWL tokens in the stimulus for all three tasks appear to significantly influence the amount of AWL tokens used in the responses. The item stimuli for these tasks could be routinely analysed through the RANGE program as part of the item development process to ascertain the percentage of AWL tokens so that the item stimuli (and therefore hopefully the responses) include at least four percent of academic tokens. Since this is a test of academic English, the higher the percentage of academic tokens elicited by each item, the stronger the claim for the test's lexical validity will be.

Another important issue here is the extent to which the items elicit the same AWL words as those in the stimulus. Repetition of AWL words in the RW-SUMM and LW-SUMM sample items appears to be an important strategy for successfully completing these summary tasks. Repetition of key words was also evident in O'Loughlin and Wigglesworth's (2007) study of test taker responses to an information transfer task. Unsurprisingly, repetition was less evident in the test of free production, WW-ESSA. Instead, the sample item analyses reveal wide use of academic words not included in the stimulus or prompt. This finding underscores the importance of including both summary and free production tasks to gain a comprehensive picture of test takers' academic vocabulary knowledge and use.

The findings also revealed a strong relationship between the global scores assigned to test taker essay responses and academic vocabulary use, in terms of tokens and, more particularly, types. Frequency and, more particularly, breadth of academic vocabulary use, therefore, appear to be important indicators of quality in the essays. It would be worth investigating whether this is also true for the automated scoring process which will be used in assessing written and spoken production on the PTE Academic in future.

Finally, while RANGE program allows for comprehensive quantitative analyses of the vocabulary used in large numbers of test taker responses, the analyses do not shed light on the extent to which the words were used appropriately in context. This is an important challenge for future work in this area.

References

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing 20*(1), 1-25.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34*(2), 213-238.

Coxhead, A. & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.) *Research perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139-155.

Foster, P. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 26, 59-84.

Heatley, A., Nation, I.S.P. and Coxhead, A. (2002) RANGE and FREQUENCY programs. http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hyland, K. & P Tse (2007) Is there an 'academic vocabulary'? *TESOL Quarterly 4*(2), 235-253.

Laufer, B. & Nation, P. (1995) Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics 16*(3), 307-322

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge: University Press.

O'Loughlin, K. & Wigglesworth, G. (2007). Investigating task design in academic writing prompt. In L. Taylor and P. Falvey (eds.), *IELTS Collected Papers: Research in speaking and writing assessment,* Cambridge: Cambridge University Press.

Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influence on foreign language performance. *Language Teaching Research, 1,* 185-211

Skehan, P. & Foster, P. (1999). The influence of task structure and processing conditions on narrative retelling. *Language Learning* 49(1), 93-120.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy and N.H. Hornberger (Eds.), *Encyclopaedia of language and education (2nd edition) Volume 7:Language testing and assessment* (pgs.111-112). New York: Springer.

Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*: 85-106

Wigglesworth, G. (1999). Rating accuracy and complexity in written scripts. Paper presented at Japanese Association of Language Teaching conference, Tokyo, October 8-10.

APPENDIX A

Task instructions, sample item stimuli and test taker responses

RW-SUMM

Task instructions:

Read the passage below and summarize it using one sentence. Type your response in the box at the bottom of the screen. You have 10 minutes to finish this task. Your response will be judged on the quality of your writing and on how well your response presents the key points in the passage**.**

Sample item stimulus (reading passage):

Just-in-Time' is a method of manufacturing products which aims to minimise production time, production costs, and the amount of stock held in the factory. Raw materials and supplies arrive at the factory as they are required, and consequently there is very little stock sitting idle at any one time. Each stage of the production process finishes just before the next stage is due to commence and therefore the lead-time is significantly reduced. With a 'just-in-time' production system, the level of production is related to the demand for the output (i.e. the number of orders) rather than simply producing finished goods and waiting for orders. This means that raw materials and stock only need to be ordered from suppliers as required.

Sample response:

Just-in-Time is a good manufacturng method which can minimize production time and cost, so that the products are manufactured depending on the demand.

LW-SUMM

Task instructions:

You will hear a short lecture. Write a summary for a fellow student who was not present at the lecture. You should write 50-70 words. You have 10 minutes to finish this task. Your response will be judged on the quality of your writing and on how well your response presents the key points presented in the lecture.

Sample item stimulus (audio-recorded listening passage):

I have chosen *The Search for Stability* as the title of my lectures, because I want to deal specifically with macroeconomics, the question of how we can keep the economy on a reasonably stable growth path. While there are disagreements about many aspects of economics, such as those dealing with efficiency, income distribution, or the role of the market versus the role of the state, I think there is widespread agreement across the political spectrum that stability is a good thing.

Economically, the first half of the 20th century was disfigured by the Great Depression of the 1930s, and the second half by the high inflation of the 1970s. No-one wants a repeat of these episodes, nor of some of the other disruptions that have marked the past 60 years. To some, the word 'stability' sounds unexciting, and probably more so if I use the term 'economic stability'. But stability is not just an economic concept; it has a profound impact on the lives of people. Instability can create havoc, damage institutions, and leave a legacy from which some families and nations will take many years to recover. For example, the rise of Nazism in Germany was helped by the preceding Weimar hyper-inflation. Fortunately, in Australia, we've had nothing like that, but the effects of the Depression left scars that lasted for lifetimes. Likewise, the effects of the big

rise in unemployment and inflation in the 1970s have not fully passed out of our economy.

Sample response:

The lecture was about economic stability. Macro economics has a reasonably stable growth path. Stability is good, although a lot of people will disagree. We should learn from our mistakes so as not to repeat the depression during the 1930's or the inflation duting the 1970's. The nazis had hyper inflation and fortunately australia was lucky enough not to have it.

WW-ESSA

Task instructions:

You will have 20 minutes to plan, write and revise an essay about the topic below. Your response will be judged on how well you develop a position, organize your ideas, present supporting details, and control the elements of standard written English. You should write 200-300 words.

Sample item stimulus (essay prompt):

Education is a critical element of the prosperity of any nation. The more educated the people in a country are, the more successful their nation becomes." Discuss the extent to which you agree or disagree with this statement. Support your point of view with reasons and/or examples from your own experience or observations.

Sample response:

Education plays a key role in the way a nation develops how ever, it is not the only element involved in the growth or success of a nation. The process of development begins with education, nations need to provide their students quality education to form competitive managers in the future. But that's only the first step in the process, the second step is to provide those students the opportunity to develop new skills in challenging works. This is the most important area or where the success of a nation really takes places. The people that work are the ones that drive the nation, hence, they are the ones responsible of the nations success. A nation becomes competitive and successful when people works had and achieves organisation's goals. Here is where another element of the process appears, motiviation at work. Companies need to provide rewards to their employees to maintain them motivated. The company's role for the success of a nation is quite important because they help the nation to achieve this success they are the ones behind this. Companies need to develop strategies to attrack, retain and motivate future and actual employees to achieve their organisational goals. When employees are motivated, their performance and productivity increases. If employees increase their productivity, the organisation is benefitiated, therefore, at the end, the nation grows. As can be seen, education is important, as the first step in the process of being a successful nation, however, it takes more than education to achieve this purpose.

APPENDIX B

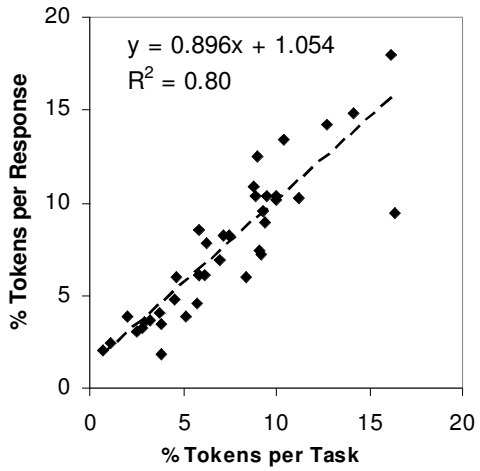

Figure 1: Percentage of AWL Tokens in RW-SUMM, stimuli and test taker responses

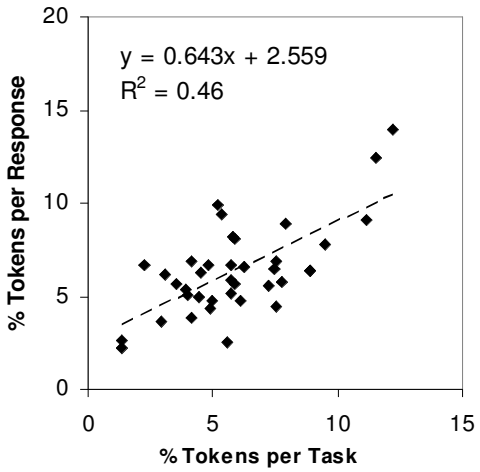

Figure 2: Percentage of AWL Tasks in LW-SUMM, stimuli and test taker responses

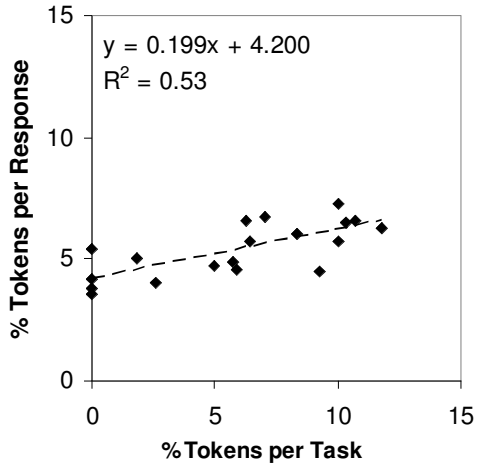

Figure 3: Percentage of AWL Tasks in WW-ESSA, stimuli and test taker responses

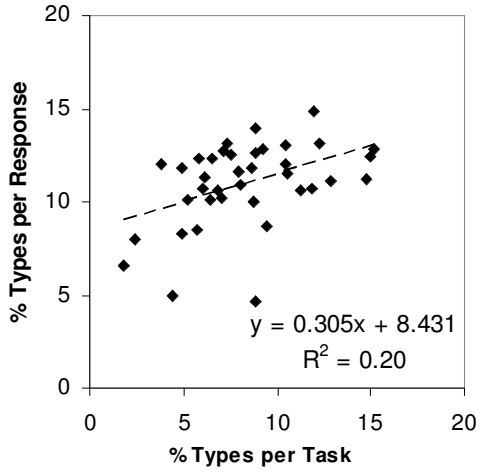

Figure 4: Percentage of AWL Types in RW-SUMM, stimuli and test taker responses

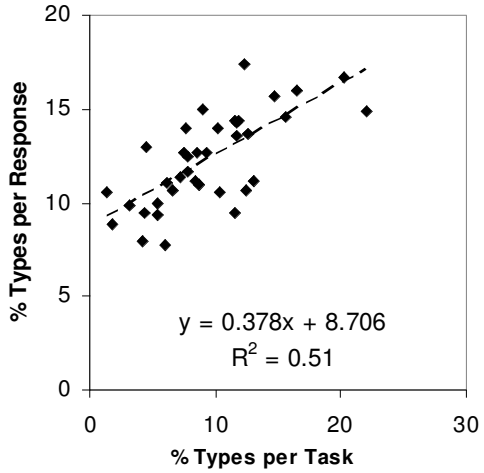

Figure 5: Percentage of AWL Types in LW-SUMM, stimuli and test taker responses

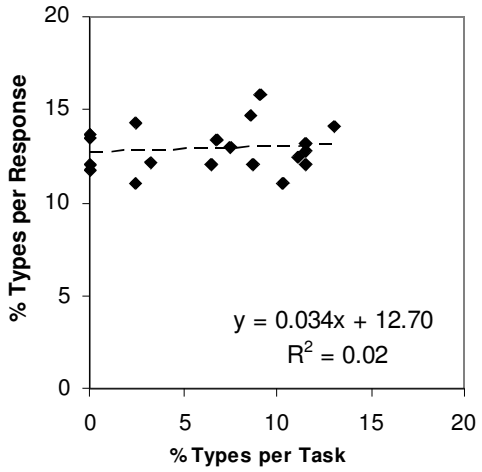

Figure 6: Percentage of AWL Types in WW-ESSA, stimuli and test taker responses