Research Note: Can PTE Academic be used as an exit test for a course of academic English?

Nathaniel Owen

University Of Leicester, UK

July 2012

1. Introduction

Universities in English-speaking countries require prospective students from those countries which do not speak English as their first language to provide evidence of their proficiency in academic English prior to the commencement of any course of study. Evidence is presented to the academic institution in the form of a test score from one of the major international test administrators (IELTS, TOEFL or Pearson). This score is then used as part of the admissions process to determine the relative success or failure of each applicant. Also prior to the commencement of any course of academic study, a course of academic English may be undertaken to further facilitate the admissions process. At the University of Leicester, such programs take place at the English Language Teaching Unit (ELTU). The University of Leicester accepts TOEFL and Pearson PTE Academic scores in addition to IELTS in an effort to recruit students from all parts of the world. As an increasing number of foreign students are now electing to undertake PTE Academic, this has been added to the list of recognized examinations.

Score equivalences between these examinations are often displayed on institutional websites, copied and distributed around the world[1]. These tables are often of unknown origin and do not represent research-based findings. Score comparison is made difficult due to differing scoring systems: IELTS band scores are reported in incremental half-band increases, whereas TOEFL and PTE scores are presented on a continuous scale. In addition, a claim of comparison across the tests may be invalid due to differing content, differing conceptions of the target domain reflected in different item types, and potentially different claims made about students who sit the different examinations. Nonetheless, as these tests aim to make similar decisions about test takers, that is, their readiness for higher education in the medium of English, questions will always remain regarding score and content comparability. Subsequently, such research has been undertaken by the Educational Testing Service (ETS) in the United States[2]. The sliding scale shows that a range of scores that are concordant with a gross band score (e.g. scores between 60 and 78 on the TOEFL iBT are reported as concordant with IELTS band 6).

Equivalence of test scores has become an increasingly urgent area of research for both academic institutions and professional bodies. Increasing demand for places at UK universities by overseas students[3] has resulted in increased pressure for English-language examinations to be valid and reliable. In addition, claims of equivalency between internal and external tests need to be verified due to the UK

[1] http://secure.vec.bc.ca/toefl-equivalency-table.cfm

[2] http://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf.

[3] http://www.ukcisa.org.uk/about/statistics_he.php#table1

government's Tier 4 legislation requiring minimum levels of competence prior to commencing academic courses[4]. The industry is now particularly competitive. Against this backdrop, individual institutions that implement their own English language examinations are also aware of these pressures. The present study is therefore a result of shared concern between the researcher, Pearson, the UKBA and the higher education community in the United Kingdom, and is in direct response to a call for research advertised on the PTE external research webpage.

2. Background

The present study focuses on PTE Academic in relation to the exit tests offered on ELTU Courses C, D and E. The ELTU at the University of Leicester gave permission for this study to be undertaken during the summer and autumn of 2011. Institutional support was therefore guaranteed by both the University of Leicester and Pearson Education Ltd. Students participating in pre-sessional Course C were invited to participate in the pilot study, and members of Courses D and E were subsequently invited to participate in the main study. The pilot study served as the basis for the researcher's master's thesis, and as the pilot study for the main research report, supplementing the larger quantitative study undertaken at the behest of Pearson and the English Language Teaching Unit.

3. Objectives, Research Questions and Hypotheses

The central research question demands that the notion of *suitability* is to be sufficiently defined and operationalized. The central thrust of the research was to determine whether or not PTE could perform the same function as the ELTU exit test, and for that to be the case, a tripartite notion of 'suitability' (figure 1) was elaborated:

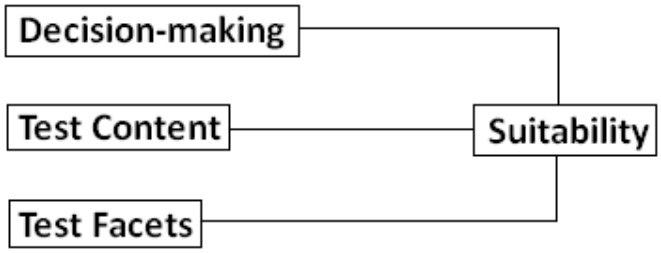

Figure 1: A three-way definition of suitability

A three-way definition of suitability highlights three key aspects that are fundamental to accurately answering the stated research question. This in turn formulates subsidiary research questions from which testable hypotheses emerged:

RQ1. Are decisions made on the basis of participants' test scores in PTE Academic and the ELTU exit test comparable?

RQ2. Are the contents of the two tests comparable?

Extending the enquiry beyond an analysis of test scores is advantageous for two principal reasons. Firstly, test performance is affected by such variables as the organization, content, format and presentation of input (electronic/paper; one-way/two-way with an interlocutor). Secondly, establishing the similarity of content across tests is an important step in the consideration of *content validity*. Bachman et al. (1995) identify a broad range of test task characteristics that may be examined in the context of a validation study. The present study aimed to build on

[4]http://www.ukba.homeoffice.gov.uk/visas-immigration/studying/adult-students/

this work.

Three expert raters were presented with example activities from PTE Academic and the ELTU exit test, alongside a 'rating-questionnaire', which incorporated a matrix of the item types from the two examinations, alongside the taxonomy of skills which are claimed to assessed by PTE Academic, as outlined in *The Official Guide to the PTE Academic* (Pearson, 2010). Additional skills as outlined in Bachman's conception of communicative language ability (CLA) (table 1) were presented to raters should they feel the Pearson taxonomy to be incomplete. The raters assigned a value from 0-4 from Bachman's five-point CLA rating instrument (Bachman et al., 1995) to measure the extent to which they felt a particular skill was involved in a particular item and communicative language ability (CLA).

Table 1: Components of Communicative Language Ability (Bachman, 1995: 189)

Language Competence			
Organizational Competence		Pragmatic Competence	
Grammatical Competence	Textual Competence	Illocutionary competence	Sociolinguistic competence
Vocabulary Morphology Syntax Phonology/ graphology	Cohesion Rhetorical organization	Ideational functions Manipulative functions Heuristic functions Imaginative functions	Sensitivity to dialect or variety Sensitivity to register Sensitivity to naturalness Cultural references and figures of speech

Bachman et al (1995) devised this five-point rating scale to be used to analyses test items for the components of communicative language ability (CLA). The rating scale incorporated considerations of whether or not the skill in question was involved, and if so, to what extent:

Table 2: Five-point CLA rating instrument (Bachman et al, 1995: 102)

Not required	Somewhat involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

The authors' technique for determining the similarity between the FCE and TOEFL tests based on this scale was employed:*" if the difference between the means for each CLA and TMF*[test method facet – see RQ3] *across the two tests was larger than the standard deviation for that facet on either of the two tests, this was interpreted as a meaningful difference"* (Bachman, 2004: 274). This technique was applied to the expert ratings. Average ratings and their corresponding standard deviations were calculated for each identified skill. Thus, the hypothesis and null hypothesis for this research question were written as follows:

Hypothesis (H_1)

$\bar{X}_1 - \bar{X}_2 \leq S_1 or S_2$

Null hypothesis $[(H]_0)$

$\bar{X}_1 - \bar{X}_2 \geqslant S_1 or S_2$

RQ3. Are the facets of the two tests comparable?

The third aspect of the research provides a quantitative approach to the notions of test content and *test method facets* (Bachman, 1990).These are "aspects of the test method that may have an impact on test scores, such as the *rubric, input or expected response*" (Fulcher & Davidson, 2007: 376. Emphasis added). Aspects of the tests that included prompt material, namely the receptive skills of reading and listening, were analyzed extensively.

Table 3 displays the categories of those facets relevant to the study, that could be readily operationalized using the vocabulary profile option of the lextutor.ca website:

Table 3: Relevant facets for analysis and method of analysis – reading and writing

Method of analysis (reading and writing prompt and output material)	Method of analysis (speaking and listening prompt and output material)
No. of items	No. of items
Average length #words	Average length #words
Vocab 1-1000	Vocab 1-1000
Vocab 1001-2000	Vocab 1001-2000
Vocab Academic	Vocab Academic
Vocab Off-list	Vocab Off-list
Flesch-Kincaid	Flesch-Kincaid
Lexical Density	Lexical Density
No. of paragraphs	(average) Length of recording #seconds
Sentences per paragraph	Fluency (words/second)
No. of words/sentence	Average number of syllables per word
Average number of sentences	Average number of syllables per second

Bachman's technique for determining the similarity between the FCE and TOEFL tests was also employed for the third research question:

Hypothesis (H_1)

$$\bar{X}_1 - \bar{X}_2 \leq S_1 or S_2$$

Null hypothesis $[(H]_0)$

$$\bar{X}_1 - \bar{X}_2 \geqslant S_1 or S_2$$

Since both tests are deemed to relate to the same target domain, the hypothesis states that the two tests will have similar content and that from the computed figures few differences will emerge. The null hypothesis states the opposite. If any differences do occur, these were flagged and analyzed – it is important to consider why these differences occur, since they may be due to an aspect of test design or a more fundamental difference in interpretation of the requirements of the target domain. Analysis centered on the exit test offered by the ELTU and three practice versions of PTE Academic included on the CD-ROM accompanying *The Official Guide to PTE Academic.*

4. Findings and Discussion

RQ2. Are the contents of the two tests comparable?

4.1 Reading tests

Analysis of the reading texts and questions yielded no statistically significant differences of opinion between the raters regarding the skills targeted by the two tests. All mean ratings fitted into the relevant SD range comfortably. However, there were two important caveats to this finding: firstly, that the means are fairly close together in five-point Likert items, suggesting any nuanced opinion of the raters was lost in a fairly crude measuring instrument. Secondly, note that for a small range of possible values the standard deviations are fairly large; in nearly all cases they were greater than one in some instances as large as the means themselves, and for several values, larger. This suggests that there was little agreement between the raters regarding the presence of some skills in particular items.

4.2 Writing tests

Similarly to the reading section, no significant differences were found between the skills on this section of the test, as determined by the ±1SD range. A similar problem emerged; that the standard deviations were in some cases larger than the means they determine to be a valid indicator of central tendency. Lower standard deviation values could be found in relation to the following three skills: 'using words and phrases appropriate to the context'; 'using correct grammar'; and 'using correct spelling'. In relation to the ELTU writing activity, there was substantial agreement between the raters on the skills tested. On these three skills, they reported an average rating of 3.57 with a standard deviation of 0.58. The median value was 4, thus indicating that for this item (there was only one item that tested writing in the ELTU test – an extended essay question) raters felt that these three skills were critical at an advanced level in order to successfully complete the item. Findings were similar on the Pearson items, standard deviations were lower than on other skills (less than 1), and median rating values were also significant – 3 and 3.5, indicating that raters felt that these skills were required at a critical intermediate level. These skills represent textual and grammatical competence in relation to Bachman's components of communicative language action, and reinforce the assertion made by Weigle (2002: 178) that writing items still prioritize the linguistic abilities of the test taker over the abilities to organize writing or produce summaries; skills consistently linked with an academic writing target domain.

4.3 Speaking Tests

Within the speaking sections of the two tests, several significant differences emerged. Table 4 summarizes these differences and the directions in which they occur (i.e. which recorded a higher mean):

Table 4: Meaningful differences in skills targeted (speaking)

Area	Direction of difference
Reading aloud	ELTU < Pearson
Supporting an opinion with details, examples and explanations	ELTU > Pearson
Using words/phrases appropriate to the context	ELTU > Pearson
Using correct grammar	ELTU > Pearson

As expected, the read-aloud component of the Pearson speaking exam scored significantly more highly than the ELTU exit test in its relevant skill, as the latter

does not include a read-aloud component. This task involves listening to a short extract of speech and then repeating what has just been heard as close to the source text as possible. This item type represents an innovative approach to the traditional dictation activity, in which test takers would be instructed to write what they had heard. This originally was deemed to be a good test of *expectancy grammar* (Oller, 1971; 1979; Oakeshott-Taylor, 1977). These items relate to phonemic and tonic production and the test taker's sensitivity to intonation. As the recording is uttered instantly after it is heard, there are few demands on a test taker's short-term memory. Similarly, the short nature of the recordings means that there is little chance of *redundancy* (Buck, 2001: 67) in the utterances, although aspects of connected speech, such as elision, intrusion or linking, may be prevalent. The extent to which a test taker is able to incorporate these within their own recording will affect their performance.

The other areas of difference, 'supporting an opinion with details examples and explanations'; 'using words/phrases appropriate to the context' and 'using correct grammar' are all rated more for the ELTU test. ELTU test activities are fundamentally different in execution to the Pearson activities; firstly, the ELTU test involves interaction with an interlocutor; secondly, one activity involves extended discussion in order to arrive at an agree conclusion; and thirdly, the prompt material is significantly different. ELTU prompt materials involve images and very few words, whereas Pearson activities may involve a diagram, graph, PowerPoint slide, text or alternatively, test takers may be required to summarize a lecture. It may be inferred that raters felt that the extended discussion of the ELTU task involved greater opportunities to support one's position with examples and greater scope for language choice due to the more generic nature of the activities. Pearson activities are more restricted in the desired output.

Nonetheless, there were several areas where there was significant agreement between the tests. 'Speaking at a natural rate' and 'producing fluent speech' were two skill areas that produced identical means, medians and very similar standard deviations. The medians for these two areas was 3, indicating a 'critical intermediate' level of engagement, indicating that all the raters agreed that these were important skills in successfully completing the speaking component of both examinations. Regardless of the mode of assessment, whether it is electronic or through an interlocutor, intelligibility is key in the success of all items, for which these skills are mandatory. The degree of agreement between raters may be backed up by considering the definition of 'fluency' in these tests, as determined through sample speaking items and the rate of speech of listening recordings, which were analyzed in relation to RQ3.

4.4 Listening Tests

Within the ±1SD range, there are no areas of difference between the tests. In terms of median values, there are nine instances of agreement, suggesting a high degree of overall agreement between the raters regarding the listening components. Regarding the standard deviations, however, on the Pearson ratings, there were no SD values below 1, whereas for the ELTU ratings there were 10. This suggests that the raters had a much higher level of agreement on the ELTU ratings than the Pearson ones and it may be concluded that the raters found PTE Academic harder to judge. Since two of the expert raters were ELTU staff, and had experience of delivering the Course C test, this is likely to have affected their level of agreement. In addition, the integrated skills aspect of the Pearson items may have made the items more difficult to judge in terms of identifying specific skills.

4.5 Inter-rater reliability

In relation to the ratings, that standard deviations throughout the expert ratings were consistently more than one, indicating a low level of agreement, it was decided

that an investigation of inter-rater reliability was required. For this, average ratings for each item type were computed for each rater across the two tests, rather than for each individual skill. Spearman's Rank Order Correlation coefficient, r_s, was selected as it is non-parametric, that is it does not predict any specific distribution (data therefore does not need to be normally distributed) and is suitable for an ordinal scale.

Raters 2 and 3 produced a statistically significant correlation of .68 ($p < 0.01$). There was also some agreement between raters 1 and 3, $r_s = .29$ ($p = 0.05$). Raters 1 and 2 produced a very weak, non-significant correlation of .26 ($p = .07$). Raters 2 and 3 were members of the ELTU, whereas rater 1 was a member of the School of Education at the University of Leicester and had no direct involvement was the creation or implementation of the ELTU end-of-course test, in contrast to raters 2 and 3. It is clear that the wide range of standard deviations meant that no significant areas of difference were discovered between the two tests, other than those areas outlined in the speaking section.

Table 5: Spearman's Rho correlations between the three raters

			Rater 1	Rater 2	Rater 3
Spearman's rho	Rater 1	Correlation Coefficient	1.000	.256	.286*
		Sig. (1-tailed)	.	.066	.046
		N	36	36	36
	Rater 2	Correlation Coefficient	.256	1.000	.683**
		Sig. (1-tailed)	.066	.	.000
		N	36	36	36
	Rater 3	Correlation Coefficient	.286*	.683**	1.000
		Sig. (1-tailed)	.046	.000	.
		N	36	36	36

*. Correlation is significant at the 0.05 level (1-tailed).

**. Correlation is significant at the 0.01 level (1-tailed).

RQ3. Are the facets of the two tests comparable?

Table 6 shows the features of the two examinations that exhibited directly observable differences and the directions of these differences.

Table 6: Direction of differences between ELTU exit test and Pearson PTE Academic and ELTU Reading & Listening texts

Area		Direction of difference
Reading	Vocab 1-1000	ELTU > Pearson
	Vocab Academic	Pearson > ELTU
	No. of sentences	ELTU > Pearson
Listening	Average length #words	ELTU > Pearson
	Vocab 1-1000	ELTU > Pearson
	Vocab Off-list	ELTU > Pearson
	(average) Length of recording #seconds	ELTU > Pearson

4.6 Reading

Pearson prompt material for the reading component displayed greater consistency in relation to the vocabulary profiles. It is also clear that the Pearson prompt material contains a greater proportion of academic vocabulary from the academic word list (AWL) (Coxhead, 2000). In contrast, the ELTU test prompts contain a greater number of sentences, which reflects that fact that PTE Academic contains a much wider variety of task types, requiring a greater variety of descriptors detailing test taker procedure and prompt material that ranges from less than 80 to more than 300. The variations in length and type undoubtedly affected the standard deviation figures and thus the determination as to whether the operationalization of the reading component displayed observable differences; the standard deviation for the length of passages for PTE Academic (70.3) reflects the fact that the 'fill in the blank' and 're-order paragraph' items were typically much shorter than the lengthy texts used for multiple-choice items. Alderson (2000) cites Engineer (1977) in arguing that longer texts may involve more discourse processing abilities rather than syntactic and lexical knowledge (Alderson, 2000: 109). The 1000 word limit suggested by Engineer that this difference may occur at is not met by any of the texts in the ELTU test or PTE Academic, however.

4.7 Listening

Transcriptions of listening materials were submitted for textual analysis in the same fashion as the reading material. Although it is accepted that reading and listening are psycholinguistically different processes, the study was concerned with a comparison of prompt material, rather than lengthy exposition detailing these different processes. ELTU test prompts had a greater proportion of vocabulary in the 'most common thousand' group; conversely, they also had more vocabulary in the off-list group, reflecting topics that were of a more specialist nature. Words that constitute the off-list proportion were mainly specialist vocabulary or culturally specific proper nouns. There were significantly fewer listening texts available to analyze than the Pearson listening recordings, and the figures may therefore have been skewed by one particular text about Bhutan, which contained vocabulary specific to that country, although reverse engineering these items determined that knowledge of this vocabulary was not deemed necessary in successfully completing the items. In addition, the ELTU recordings were longer than those of PTE Academic, which takes into account once more the greater variety of task types in PTE Academic. Dictation activities contained recordings that lasted only a few seconds, and each test contained three of these items, which greatly affected the average length of the recordings.

4.8 Speaking

The Pearson speaking and writing section consists of 39-41 questions, which vary considerably in scope. Some questions are based around a written text, others around an authentic extract of recorded speech. These extracts may be accompanied by a diagram, map, table, graph or image. In contrast, the Course C ELTU test consists of two questions, which are fixed in scope. The first question is an independent long turn item, based on that offered by IELTS, and centers on a visual prompt and a short text, such as a question that directs the student thought process towards the particular topic that is required for assessment. The extent to which the student addresses various aspects related to the image(s) is one mode of assessment. The number of students in each discussion should be a minimum of three, and the discussion is self-directed.

Table 7: Pearson PTE Academic Speaking - Analysis of sample texts (prompt material)

	Item type 1: read aloud*	Item type 2: repeat sentence	Item type 4: re-tell lecture (prompts)	Item type 5: answer short question	Average	SD
No. of items	21	34	9	35	24.75	12.28
Average length #words	58	10.41	161.56	14.91	61.22	70.25
Vocab 1-1000	82.18	80.67	80.4	83.33	81.65	1.37
Vocab 1001-2000	5.84	4.76	3.85	5.17	4.91	0.83
Vocab Academic	5.07	7.84	6.74	4.02	5.92	1.7
Vocab Off-list	6.91	6.72	9.01	7.47	7.53	1.04
Flesch Reading Ease	45.44	62.48	49.51	73.1	57.63	12.61
Lexical Density	0.54	0.55	0.56	0.51	0.54	0.02
(average) Length of recording #seconds	37.1	6.51	67.33	5.04	29	29.52
Fluency (words/second)	1.6	1.6	2.4	2.96	1.91	0.99
Average no. of syllables per word	1.64	1.58	1.63	1.46	1.58	0.08
Average number of syllables per second	2.56	2.53	3.91	4.31	3.33	0.92

*12 items at 35 seconds and 9 at 40 seconds

The Pearson and ELTU speaking tests therefore differ substantially in their goals, task orientation and setting: three aspects of a framework for describing tasks presented by Fulcher (2003: 57). Fulcher describes activities similar to repeat sentence and read aloud items as having no specific goals and being non-interactional as "there is no 'communication' taking place"(ibid.: 58). In

contrast, ELTU items are specifically goal-oriented in that they require several participants to negotiate and construct discourse to arrive at a specific conclusion on a generic topic. These are thus examples of 'performance assessment' (McNamara, 1996). Also present are two interlocutors that were also course tutors; thereby providing a level of familiarity unavailable in PTE Academic. The number of these one-directional and non-communicative items present in PTE Academic is significant; they are considered to be a valid and reliable method of assessing certain aspects of a test taker's linguistic ability; namely those related to pronunciation, intonation and rhythm. They are quick, easy and simple to implement, and computer programs – by developing a battery of responses – can readily equate responses to particular levels. From table 7 above, the nature of the input that is designed to elicit desired responses may be elucidated. Specifically, the vocabulary profile is significantly different from those of the input to the reading and listening items. Flesch reading grades, computed on transcribed versions of the prompt material, display figures of 62.48 and 73.1 for those items that utilize input of only a few words or a single sentence. Fluency ratings, however, are more consistent with those found in the listening input material.

4.9 Writing – Pearson and ELTU Response Models

For the ELTU test, test takers are expected to produce an essay-style piece of writing commensurate with the conventions of an extended essay typical of academic study. They should utilize a variety of functional and socio-linguistic features of language to produce a text that attempts to explain, evaluate or persuade an audience with the veracity of the argument in response to a specific prompt. Responses should be divided into paragraphs, with the correct linguistic features of an introduction, body paragraphs and conclusion. Lexical and grammatical features should all be commensurate with an academic, formal style of writing.

For PTE Academic, test takers are expected to summarize a text presented to them by selecting the most salient information and presenting it in their own words in a succinct manner. Success on this item is judged according to the cohesion and coherence of their response, combined with spelling and grammar. Test takers are also judged on their ability to select the most important information. Test takers are then expected to produce a lengthy piece of writing within the parameters of a stated essay title. They must use their general knowledge, as the essay topics should not discriminate on the basis of schematic knowledge– and their knowledge of academic writing conventions – to formulate an appropriate argument that explicitly answers the stated question. The success of this item is judged on the test taker's ability to formulate an argument and present it using appropriate grammar, vocabulary, spelling and presentation conventions.

Table 8: Comparison of facets across work products and sample responses

Pearson PTE Writing Sample work products					ELTU Sample work products					
	Average	SD	SD +1	SD -1	Average	SD	SD +1	SD -1	Difference	Direction
No. of items	3	0	3	3	6	0	6	6	N/A	N/A
Average length #words	295	22.61	317.61	272.39	263.83	31.96	295.79	231.34	none	N/A
Vocab 1-1000	77.06	2.91	79.97	74.15	80.84	3.04	83.88	77.8	difference	ELTU > Pearson
Vocab 1001-2000	6.77	2.83	9.6	3.94	5.7	3.13	8.83	2.57	none	N/A
Vocab Academic	9.54	3.46	13	6.08	10.08	4	14.08	6.08	none	N/A
Vocab Off-list	6.62	5.86	12.48	0.76	3.37	0.56	3.93	2.81	none	N/A
Flesch Reading Ease	43.23	0.01	43.24	43.21	48.86	8.49	57.35	40.37	none	N/A
Lexical Density	0.56	0.01	0.57	0.55	0.53	0.01	0.54	0.52	difference	Pearson > ELTU
Average number of paragraphs	4.33	0.3	4.63	4.03	3.83	0.75	4.38	3.08	none	N/A
Sentences per paragraph	3.28	0.3	3.58	2.98	4.06	0.99	5.05	3.07	none	N/A
No. of words/sentence	22.49	4.06	26.55	18.43	17.77	2.48	20.25	15.29	difference	Pearson > ELTU
No. of sentences	13.33	1.53	14.86	11.8	15	1.67	16.67	13.33	none	N/A

Pearson work products were included on the CD-ROM. ELTU work products were samples of text from participants who had successfully completed the pre-sessional course and whose written work was deemed of sufficient quality to commence their academic program. The work products and sample items across the ELTU exit test and Pearson PTE are remarkably similar on the basis of the above analysis. The work products of the research participants had a higher proportion of vocabulary within the most common thousand. Two other differences emerged: Pearson sample responses had a slightly higher lexical density, and a higher number of sentences overall. Neither of these differences was deemed significant in relation to the overall aims of the item types, and therefore it is reasonable to suggest that similar decisions regarding these work products would have been made.

5. Conclusions

Is PTE Academic suitable as an exit test for a program of academic English?

Based on the available evidence from this research project, the above question may be answered in the affirmative.

Regarding the *test contents*, although it was clear that the ratings provided by the

expert raters contained some ambiguities (large standard deviations suggested that the spread of ratings around each mean was wide, and that therefore there was little agreement regarding the contents of the tests), some significant differences were posited regarding the speaking aspects of the tests. These differences were due to the different conceptions of the construct of 'academic speaking' inherent to the items, as determined by the raters. However, data presented regarding the elicited imitation (EI), or read aloud, items suggest that these items are capable of recording reliability equal to that of more conversational item types (Pearson, 2011; Van Moere, 2012). As stated in the methodology section, the aim was not to critique any specific conception of the construct of academic English, but to identify differences between items.

Regarding the facets of the test, the majority of the prompt material and work products that were analyzed overlapped. Few significant differences emerged, especially regarding the work products, suggesting similar conceptions of language ability, based primarily on lexical and syntactic knowledge. Despite differences in the test items, differences were insufficient to state that this represented a fundamental difference in either of the tests' conception of academic English. Differences that fell outside of the ±1SD range tended to be related to vocabulary profiles and were minor differences; the Pearson reading test prompts had slightly more 'academic' vocabulary in comparison to the ELTU exit test and the ELTU more examples from the most common thousand group.

5.1 Limitations to the research

Unfortunately no speaking products were available for analysis. Transcription of Pearson speaking items revealed aspects of connected speech and operational definitions of fluency, but these could not be compared to work products for the ELTU exit test. Gaining access to these at a future date with future research participants could prove illuminating, especially when comparing scores that were awarded for these work products. Furthermore, opportunities for interviews with both research participants and expert raters were excluded due to time constraints. Interviews with expert raters might have afforded valuable insight into how the skills outlined in the domain analysis were interpreted by the raters. Incorporating an additional activity into the ratings involving raters assigning the skills from the Pearson rubric into Bachman's conception of CLA, and then comparing their interpretations would have leant weight to the ratings themselves. The problem with this approach is that the rating activity is already substantial. Increasing the volume of work to be undertaken by raters may deter some from participating, regardless of remuneration offered. Similarly, interviews with test takers to determine which test items they had difficulty with and why they had difficulty would have given more credence to the test scores, as well as providing insight into which items the test takers considered were more appropriate, given the content of the course they had just undertaken.

5.2 Future Research

Overall, while the content analysis and the test method facet analysis in the present study has revealed some useful information regarding the make-up of the two tests, insufficient analysis of work products and the numbers of test items in relation to the ELTU exit test has partially undermined the findings. A greater focus on specific components of the exams (speaking, reading, writing or listening) is needed to answer the research question with more confidence for each component. A greater number of test items also need to be analyzed in order to make more determined conclusions, especially for the ELTU exam.

In addition, a 'third layer' is proposed for any future study of this nature. Once research participants have completed a semester in their chosen program, researchers could contact their personal tutors and posit several direct 'yes/no'

questions such as:

> "Is the student's English language proficiency sufficient for them to successfully complete their academic program?"
>
> "Remedial English language courses are available through the ELTU. In your opinion, does the student require such a program to successfully complete their program?"

Such questions would elicit practical responses from these members of staff, who have dealt with the research participants in both an academic and pastoral role and are therefore well-placed to make such subjective judgments. The direct nature of these questions also renders them amenable to statistical analysis, providing another means of comparison across test scores. The responses would assist in determining which of the examinations had better predictive validity in terms of assessing students' suitability to undertake a course of academic English. Specifically, because of the nature of the relationship between personal tutors and students, these tutors would be well-placed to make evaluative judgments in relation to their speaking and writing abilities, which are more difficult to compare statistically. Different results between the subjective reports of the tutors in comparison to the test scores would provide more insight into these areas.

References

Alderson, J. C. (2000) *Assessing Reading*. Cambridge: Cambridge University Press.

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K. & Choi, I. C. (1995) *An investigation into the comparability of two tests of English as a foreign language: The Cambridge TOEFL Comparability Study.* Cambridge: Cambridge University Press.

Bachman, L. F. (2004) *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Buck, G. (2001) *Assessing Listening*. Cambridge: Cambridge University Press.

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

Davidson, F. & Lynch, B. K. (2002) *Testcraft: a teacher's guide to writing and using language test specifications.* New Haven, CT: Yale University Press.

Engineer, W. (1977) *Proficiency in reading English as a second language.* Unpublished PhD thesis, University of Edinburgh.

Fulcher, G. (2003) *Testing second language speaking. London*: Longman/Pearson Education.

Fulcher, G. and Davidson, F. (2007) *Language testing and assessment: an advanced resource book* (Routledge Applied Linguistics Series). Oxford: Routledge.

Oakeshott-Taylor, J. (1977) *Information redundancy and listening comprehension*, in Dirven, R. (ed). Listening Comprehension in Foreign Language Teaching. Kronberg: Scriptor.

Oller, J. W., Jr. (1971) *Coding information in natural languages*. The Hague: Mouton.

Oller, J. W., Jr. (1979) *Language tests at school: a pragmatic approach*. London: Longman.

Weigle, S. C. (2002) *Assessing writing.* Cambridge: Cambridge University Press.

Test Material

The Official Guide to PTE Pearson Test of English Academic (2010). Pearson Education Asia Limited.

Web Resources

http://www2.le.ac.uk/offices/eltu/presessional

http://secure.vec.bc.ca/toefl-equivalency-table.cfm

http://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf.

http://www.ukcisa.org.uk/about/statistics_he.php#table1

http://www.ukba.homeoffice.gov.uk/visas-immigration/studying/adult-students/

http://www.ukba.homeoffice.gov.uk/sitecontent/newsfragments/72-LMU-student-page

http://www.telegraph.co.uk/education/universityeducation/9497191/Universities-admitting-foreign-students-with-poor-English.html