

Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic)

Mehdi Riazi

Department of Linguistics, Macquarie University

This study examines the concurrent and predictive validity of the newly developed Pearson Test of English Academic (PTE Academic). The study involved 60 international university students who were non-native English speakers. The collected data included: the participants' scores on a criterion test (IELTS Academic), their PTE Academic scores, and their academic performance as measured by their grade point average (GPA). The academic performance data of a similar norm group of native speakers were also collected. Results of the data analysis showed that there is a moderate to high significant correlation between PTE Academic and IELTS Academic overall, and also in terms of the four communication skills of listening, reading, speaking, and writing. Additionally, significant correlations were observed between the participants' PTE Academic scores (overall and the four communication skills) and their academic performance. Results show that as the participants' PTE Academic scores increased, their academic performance became on par or exceeded that of the norm group such that those in C1 and higher levels of the Common European Frame of Reference (CEFR) outperformed the norm group academically. Findings of this study provide useful implications for the testing community and higher education decision-makers.

Key words: validity, criterion-related validity, PTE Academic, IELTS Academic

Introduction

This study fills the gap in the literature by reporting a validity study of the newly launched PTE Academic. The study bears significance given that Australian universities and institutes of higher education have started to

¹ Address for correspondence: A/Prof. Mehdi Riazi, Department of Linguistics, Macquarie University, North Ryde, NSW 2109, Australia; Email: mehdi.riazi@mq.edu.au.

recognise PTE Academic along with IELTS Academic and TOEFL-iBT; and this trend is likely to spread more widely across Australian universities and institutes of higher education in the future. Accordingly, it is imperative for different stakeholders to have more empirical evidence at their disposal related to this test in order to be better informed in their decision-making. Moreover, this study contributes to our knowledge base of test validation by focusing on a newly launched large-scale and high-stakes test. While there are quite a lot of validity studies on the two more established language proficiency tests (IELTS and TOEFL), there is a scarcity of such studies for PTE Academic after its formal launch in 2009. This, indeed, may be due to the fact that the test is new. We will hopefully see more empirical studies on the validation of this test which will contribute to our knowledge base regarding this English language proficiency test.

The study reported here aims to investigate concurrent and predictive validity of PTE Academic. Concurrent and predictive validity are essentially two types of criterion-related validity (Hughes, 2003); the former seeks correlations between two tests (the test and the criterion), which are almost simultaneously administered, and the latter concerns the extent to which a language test can predict the future (academic) performance of test-takers. Since IELTS Academic is by default the test recognised by Australian higher education institutions, it was used as the criterion in the present study. Before the current study is presented, a review of the related literature will help to put the study in perspective.

Literature Review

The influx of international students into English speaking countries to pursue their tertiary studies is on the rise. Therefore, matriculation regulations of universities and institutions of higher education in Anglophone and other countries where the medium of instruction is English require international non-native students to fulfil the English language entry requirement. Apart from a few locally and in-house developed English language proficiency tests (see, for example, Malone, 2010; Zhang, 2009), currently the International English Language Testing System (IELTS Academic), and Test of English as a Foreign Language (TOEFL) Internet-based Test (iBT) are the two dominant proficiency tests, which are taken by international students to fulfil tertiary institutions' language entry requirements. Most universities recognise both TOEFL and IELTS. However, TOEFL is more commonly taken by prospective students who intend to pursue their studies in North American universities, whereas IELTS is more popular in European and Asia Pacific countries and universities. Yet,

IELTS Academic is by default the test recognised by Australian higher education institutions and is the dominant test taken by prospective international students who apply to Australian universities. Both IELTS and TOEFL-iBT assess the test-takers' English language ability in the four skills of listening, reading, speaking, and writing. IELTS Academic reports test-takers' scores using a band score of 0–9 with half increments for individual skills and an overall score, while TOEFL-iBT reports test-takers' results of each skill on a scale of 0–30 and an overall maximum score of 120. For a review of the IELTS Academic writing module and TOEFL-iBT, readers can refer to Uysal (2010) and Alderson (2009) respectively. Further, the scope of this study does not allow a review of the predictive and concurrent validity studies of the criterion test (IELTS Academic) here. Readers may refer to recent studies such as Ingram and Bayliss (2007).

Another international English language proficiency test, which was launched globally in 2009 with the same purpose as IELTS Academic and TOEFL-iBT, is the Pearson Test of English Academic (PTE Academic). It is a computer-based international English language test, which is designed to assess English language competence in academic contexts (Zheng & De Jong, 2011) through measuring the non-native speakers' abilities in reading, writing, listening, and speaking. PTE Academic reports results using a band score of 10–90 both overall and for individual skills. The report also provides scores for enabling skills (grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse) although these scores are only provided for information and not included in the overall scores as well as each skills score (Pearson, 2010).

Examining and providing evidence for the validity of large-scale and high-stakes English language proficiency tests is one of the major concerns of both test developers and test score users. Test developers may modify their tests and test score users may make better decisions using the evidence provided through empirical studies. Based on the researcher's review, currently, the evidence for the validity of PTE Academic is predominantly reports from the test organisation which developed the test. These reports are based on the studies conducted during the development of the test and by collecting and analysing the data from the field test. These studies can be divided into two groups. First, reports from the test organisation, and second, reports from independent researchers whilst still relying on the data from the field test. Two main reports are provided by the test organisation. The first one is a benchmark study (Pearson, 2010) in which data were collected from the first stages of developing the test and benchmarking it to the Council of Europe Frame of Reference (CEFR) (Council of Europe, 2001). Since the CEFR was developed, its levels have been used for benchmarking language learners' ability both in terms of

pedagogy and assessment and in different contexts (see, for example, Mader & Urkun, 2010). It is thought CEFR enables different stakeholders, including language learners, teachers, universities or potential employers, to use CEFR for the purpose of comparing and relating language proficiency certificates by different levels of CEFR. Pearson's report of aligning PTE Academic with different levels of CEFR is thus aimed at this target. The second report includes two rounds of field tests (August–October 2007) and (May–June 2008) in which both native and non-native test-takers took part. Psychometric analysis of PTE Academic data were performed, and concordance studies of PTE Academic scores with other major English language tests, i.e., TOEFL and IELTS scores were accomplished (Zheng & De Jong, 2011).

Of the above two reports, Zheng and De Jong (2011) is related to the present study. They provide two sources of evidence for the concurrent validity of PTE Academic. The first includes the statistical validation procedures used to establish the extent to which PTE Academic scores can be linked to the CEFR. The second data source they report presents the results from a concordance study between PTE Academic and other measures of English language competencies during the field-testing stage. They used test-takers' self-reported scores for the other tests of English, including TOEIC, TOEFL-PBT, TOEFL-CBT, TOEFL-iBT, and IELTS. About one in four of all test-takers (13 for the TOEFL-iBT group and 15 for the IELTS group) who provided self-reported scores also sent in their official test score reports, which were used to check the reliability of the self-reported scores ($r = 0.82$ for TOEFL-iBT and 0.89 for IELTS). They then used Educational Testing Service (2005) guidelines to convert their participants' TOEFL-CBT scores to TOEFL-iBT and TOEIC scores to TOEFL-PBT. Afterwards, the researchers used regression coefficients to predict test-takers' scores on TOEFL-iBT and IELTS using PTE Academic BETA² scores (Zheng & De Jong, 2011).

Ultimately, based on the results from the concordance of PTE Academic with CEF and other English tests, two concordance tables are generated for PTE Academic scores with CEF and TOEFL-iBT, and CEF and IELTS Academic scores respectively.

While Zheng and De Jong (2011) provided useful information about the construct and concurrent validity of PTE Academic, two issues need to be addressed here. First, their study was based on the data collected from the development stage of the PTE Academic test or the BETA stage. Second, the data source used by them to provide evidence for concurrent validity were

² Beta refers to the version of the test which was used in the field-testing stage

based on the concordance between test-takers' PTE Academic and their self-reported and predicted scores which, as described above, were calculated using statistical procedures. Nor is it clear from their report when the test-takers took other tests which were included in the concurrent study. Furthermore, the multitude of score conversions and the predictions they made through statistical analysis (as included in their reported correlations) might be of further concern to language testing experts because these predicted scores reached through statistical procedures and may not represent real data (scores) when participants actually take the test. Accordingly, the present study attempts to address these issues by: (a) using data (PTE Academic scores) from the real test administration and not from the BETA (field-testing) stage, (b) collecting official transcripts of the participants' IELTS Academic scores, and (c) limiting the gap between test-takers' sitting for PTE Academic and their IELTS test to a maximum of six months for studying the concurrent validity of the test.

In addition to the above reports produced by the test organisation, there are also other research projects and reports funded by the test organisation but conducted by other researchers outside the test organisation. The first of these is a research project on the lexical characteristics of PTE Academic (Pearson, 2008). This study investigated the relationship between lexico-grammatical aspects of speech, language proficiency, task type, and language background of the oral component. The study aimed at examining the lexical validity of the PTE Academic oral module. That is, the correspondence between the words occurring in, and elicited by, the PTE Academic oral module and those test-takers will encounter and are expected to produce in their real-life academic contexts when engaged in oral communication. The study also reports a correlation between the frequency of academic level word usage and the CEFR level assigned to test-takers by human raters. There are other research notes provided by Pearson which are not directly related to the present study so interested readers are urged to refer to the Pearson website (<http://pearsonpte.com/research/Pages/ResearchSummaries.aspx>) to check the research summaries and notes.

There are a few other reports, while not directly relevant to the focus of the current study, that have examined one aspect of the validity of PTE Academic. These include a master's and a PhD thesis (Chan, 2011; Lu, 2012), a review of PTE Academic (Wang, Choi, Schmidgall & Bachman, 2012), and a psychometric measurement model for adult English language learners using PTE Academic (Pae, 2012).

Chan (2011) examined the cognitive validity (cognitive processes and writing strategies) of two item types of the writing section of PTE Academic which

include: summarising written text and writing an essay. Chan used Weir's (2005) socio-cognitive framework for test validation and investigated and compared the cognitive processes of a group of ESL test-takers while they performed the integrated task (summarising written text) and the independent task (writing an essay). Chan also studied test-takers' opinions regarding the two different writing tasks, namely, the integrated and the independent. Results showed that the two writing tasks elicited different cognitive processes. The integrated task (summarising written text) was found to elicit macro-planning and discourse synthesis processes such as selecting relevant ideas from source text, whereas the independent task (writing an essay) elicited lower levels of micro-planning, monitoring and revising. In terms of test performance, Chan's results showed significantly better scores on the independent task (writing an essay) than on the integrated task (summarising written text). In terms of test-takers' views (face validity), test-takers were found to prefer an integrated task for teaching and learning, and writing an essay for assessment purposes. As indicated above, Chan's study is not directly relevant to the focus of the current study, however, it contributes to the body of knowledge of test validation in general, and to the validation of the PTE Academic writing module in particular.

On the other hand, Lu (2012) developed and applied an argument framework to assess the validity of the PTE Academic writing task scores which are derived from an automated scoring system called the Intelligent Essay Assessor (IEA). Both human raters and IEA scoring of the essays produced by PTE Academic test-takers were used to investigate the validity of scores produced by IEA. Using a wide range of empirical evidence and theoretical rationale, Lu made a number of recommendations with a view to further strengthening the validity of scores produced by the IEA. While Lu's study concentrated on the validity of the scoring system of the writing module of PTE Academic, Pae (2012) applied Rasch modelling to examine the psychometric properties of the whole PTE Academic scores. Pae used 140 test-takers' scores from the PTE Academic database to study the conformity of the test-takers' performance on the 86 items of form one of the PTE Academic field test. Based on the findings that no significant differential item functioning was found across subgroups of gender and spoken-language context, Pae concluded that PTE Academic is a valid and useful measurement tool for English language learners' academic English assessment.

Finally, Wang et al. (2012) used an argument-based approach to test validation based on Bachman's (2005) and Bachman and Palmer's (2010) validity framework and reviewed PTA Academic to investigate the extent to which the intended uses of the test can be justified to stakeholders. As is the case in an

argument-based approach to test validation, they firstly present an assessment use argument (AUA) in which explicit claims that link test-takers' performance to the consequences of test use are articulated, and secondly they provide evidence provided by the test developer to examine the claims that were articulated in the AUA. Based on their review, Wang et al. (2012) conclude that positive evidence has been identified by the test developer for most of the articulated warrants showing the primary use of PTE Academic for making admission decisions at tertiary level institutions and organisations. Rebuttal evidence was not observed from the provided documents; however, potential rebuttals were frequently noted. Accordingly, Wang et al. (2012) conclude that the evidence gathered from the review documents was relatively extensive for the claims about the assessment records and score interpretations, while they found much less evidence for the claims about decisions and consequences. They argue that perhaps this is because providing evidence for these two claims (decision and consequences) are primarily the responsibility of the local decision makers. They therefore suggest that the test developer conducts or sponsors studies to address the potential rebuttals they have noted.

The current study aims at furthering validation studies of PTE Academic especially by using data from the post field-testing stage and when the test is in wide-range use. The study can provide useful information to different stakeholders especially because, as can be seen from the above review, there are not very many empirical studies which address validity issues of PTE Academic and particularly the predictive validity of the test. Accordingly, the current study intends to fill this gap and provide some empirical evidence as relates to the criterion-related (concurrent and predictive) validity of PTE Academic.

Purpose and Research Questions

Based on the review presented above, this study pursues two main goals: the first goal is to study the concurrent validity of PTE Academic, and the second goal is to study its predictive validity. These two general goals are translated into the following research questions:

1. Is there any relationship between the participants' PTE Academic and IELTS Academic scores in terms of their overall scores and on individual scores of listening, reading, speaking, and writing?
2. Is PTE Academic able to differentiate between the two proficiency groups (higher and lower) as formed by participants' IELTS Academic scores?

3. Is there any relationship between participants' PTE Academic scores and their academic achievement as measured by their first semester grade point average (GPA) and as compared to a native-speaker norm group?
 - a. Is there any significant relationship between participants' PTE Academic scores and their first semester GPA in terms of their overall scores and the four communication skills scores?
 - b. What percentage of participants with PTE Academic scores in the A2, B1, B2, and C1 and higher range of the CEFR levels reach academic norm as compared to the norm group's GPA?

Concerning the third research question (sub-question b), it should be explained that given that an aim of the study is to investigate the relationship between participants' English language proficiency as measured by their performance on PTE Academic and their academic performance, it is useful to be able to compare the participants' academic performance with that of a group for whom it can be assumed that language proficiency (or, at least, the condition of not having English as their L1) is not a factor. This is based on the assumption that academic performance is influenced by all sorts of other factors, and so it will be interesting to compare test-takers' academic performance with a norm group for whom English language proficiency has not been an issue. Accordingly, it was thought that if a norm group of native speakers could be formed, the comparison between the participants' academic performance and that of the norm group will be revealing in light of the participants' level of language proficiency. Details of forming the norm group along with other aspects of the study are presented in the next section of the paper.

Method

Participants

Participants of this study comprised 60 non-native international undergraduate ($n = 37$) and postgraduate ($n = 23$) students at Macquarie University in Australia. Both female ($n = 45$) and male ($n = 15$) students of various nationalities (see Appendix 1) were recruited from three major academic disciplines: Sciences ($n = 10$), Business and Economics ($n = 40$), and Arts and Humanities ($n = 10$). Their ages ranged from 18–36 years (mean = 23.05; SD = 3.38). These participants had already taken IELTS Academic as a requirement for their entry into university. Those volunteer students who had taken the test at most six months prior to their taking of PTE Academic were selected. This criterion was set to facilitate data collection as it was not feasible to recruit all participants who had taken IELTS just prior to sitting for PTE Academic. The

time range between taking PTE Academic and IELTS Academic was two to six months with a mean of 4.7 months. Thirty-three (55%) participants had taken IELTS Academic between two to five months before taking PTE Academic, and 27 (45%) had taken IELTS Academic six months before taking PTE Academic. However, all participants stated in their demographic questionnaire they had not attended any English language courses after they took IELTS Academic.

The participants were recruited from the pool of students enrolled in the second semester of 2011 and the first semester of 2012, and they received an honorarium to compensate for the time they contributed to the study as well as the cost of the PTE Academic test.

Most participants were Chinese students (21 participants; 35%) followed by Vietnamese (8 participants; 13.3%), and Korean (4 participants; 6.7%). Appendix 1 presents details of the number of participants and their nationalities. Appendix 2 shows the details of the number of participants from each specific field in each discipline.

Data Collection and Analysis Procedures

The volunteer participants took part in an introductory session in which the purpose of the research was explained to them. Then they signed a consent form, completed a brief demographic questionnaire, and provided the researcher with a copy of their IELTS Academic transcript. They were then invited to another introductory session, which familiarised them with the PTE Academic test. They were instructed to complete an online registration for a real PTE Academic test using the voucher provided to them for this purpose. They also gave a copy of their academic transcript at the end of the academic semester to which they were enrolled. After the academic transcripts were collected from all the participants, their departments and the units they had enrolled in were identified. A list was sent to Macquarie University registrar's office to receive the native-speakers (Australian by birth) marks for the same units. Australia being a high profile country in terms of immigrants, the 'Australian by birth' criterion was used to select the norm group. The registrar's office was asked to identify the native speaker (Australian born) students who were enrolled in the same units as our international participants. This yielded accurate data for calculating the norm group's GPA for semester two, 2011 and semester one, 2012.

Using the median of the participants' IELTS scores, they were divided into two proficiency groups: higher and lower. The median scores were 7.5, 7.5, 6.5, 7.25, and 7.25 for the listening, reading, speaking, writing, and overall scores

respectively. In order to form the two groups for each skill and for the overall score, those participants whose scores were exactly the same as the median score, were excluded. Accordingly, except for the overall score where we ended up with exactly 30 participants in each group, the number of participants in the lower and higher proficiency groups for the four skills differed slightly as can be seen in Table 5 in the results section.

In terms of data analysis, Pearson correlation was used to answer the first research question, which seeks possible relationships between participants' IELTS Academic and PTE Academic scores. Moreover, an independent t-test was used to investigate whether PTE Academic was able to differentiate between the two proficiency groups as was formed by using the median of the IELTS scores to answer the second research question. Finally, we used Pearson correlation to answer the first part of the third research question, which looks for possible relationships between participants' PTE Academic scores and their academic achievement. In order to answer the second part of the third research question, participants' PTE Academic scores were classified according to CEFR levels (A2, B1, B2, and C1 and above) based on Pearson (2012). This categorisation was then used to calculate the percentages of the participants in each level who reached or exceeded the norm group's GPA.

To calculate the GPA for the norm group, the units in which each participant of the study was enrolled were identified. The native speakers enrolled in the same units were identified too so that their scores could be obtained from the registrar's office. Then native speakers' scores were used to calculate a corresponding norm group GPA for each participant. For example, if one participant had enrolled in three 3-point credit units of 'techniques and elements of finance', 'microeconomics principles', and 'marketing fundamentals' with the marks of 78, 61, and 68, the participant's GPA was calculated to be 69. Then, native speakers' marks were collected for the same three units and were averaged to calculate a corresponding norm group GPA for the above participant. This yielded a norm group GPA for each participant. Moreover, participants' PTE scores and their categorisations into CEFR levels made it possible to identify the participants' GPA and their corresponding norm group's GPA in each of the CEFR levels.

Results

Results of the data analysis are presented in this section in response to each of the research questions. The results will be then discussed in the Discussion section.

RQ1: *Is there any relationship between the participants' PTE Academic and IELTS Academic scores in terms of their overall score and on individual scores of listening, reading, speaking, and writing?*

Table 1 presents the descriptive statistics of the participants' scores on IELTS Academic and PTE Academic.

Table 1. Descriptive statistics for participants' IELTS Academic and PTE Academic test scores

		Min.	Max.	Mean	SD	Median	Mode
IELTS Academic	Listening	4.5	9	7.43	1.15	7.5	8.5
	Reading	4.5	9	7.52	1.11	7.5	8.5
	Speaking	4.5	8.5	6.76	0.90	6.5	6
	Writing	5	8	6.43	0.68	7.25	6
	Total	5.5	8.5	7.12	0.82	7.25	7.5
PTE Academic	Listening	37	90	64.65	14.66	64.5	71 ^a
	Reading	30	84	60.33	14.73	63	68
	Speaking	10	90	66.92	17.10	67	90
	Writing	33	88	60.5	13.27	62	56 ^a
	Total	36	90	63.42	14.10	64	50

^aMultiple modes exist. The smallest value is shown.

As can be seen from Table 3, this cohort of participants had a mean score of 7.12 and 63.42 for their overall performance on IELTS Academic and PTE Academic respectively. While they achieved their highest score (mean = 7.52) from the reading section of the IELTS, they had their highest score (mean = 66.92) from the speaking section in the PTE.

In order to determine if there is any relationship between participants' scores, overall and for each of the four communication skills, Pearson correlation was run between the participants' scores in the two tests. Table 2 presents the results of the correlation.

Table 2. Relationship between PTE Academic and IELTS Academic scores ($n = 60$)

	IELTS Listening	IELTS Reading	IELTS Speaking	IELTS Writing	IELTS Total Score
PTE Listening	0.661**				
PTE Reading		0.677**			
PTE Speaking			0.723**		
PTE Writing				0.686**	
PTE Total Score					0.824**

**Correlation is significant at the 0.01 level (2-tailed)

Table 2 shows that there is significant moderate to high correlation between participants' PTE Academic and IELTS Academic scores. The highest correlation ($r = 0.824$) was observed between participants' overall scores in the two tests showing a common variance of almost 68%. The next highest correlation was observed between participants' speaking ($r = 0.723$) and writing ($r = 0.686$) skills of the two tests.

RQ2: *Is PTE Academic able to differentiate between the two proficiency groups (higher and lower) as formed by participants' IELTS Academic scores?*

To answer this research question, the median of the IELTS Academic scores of the individual skills and the overall score as presented in Table 3 was used to divide the participants into two groups of higher and lower proficiency. In order to form two distinct groups those participants whose scores were exactly the same as the median scores were excluded because they could not be placed in one of the two groups. Accordingly, except for the overall score where we ended up with exactly 30 participants in each group, the number of participants in the lower and higher proficiency groups for the four communication skills differed slightly as can be seen in Table 3. In order to find out if PTE Academic was able to differentiate between the two language proficiency groups an independent t-test was run with the results presented in Table 3.

Table 3. Results of the independent t-test for the two proficiency groups

PTE Academic	Group	N	Mean	SD	t	df	Eta ²
Listening	Lower	25	55.56	13.31	- 5.51**	50	0.38
	Higher	27	74.26	11.13			
Reading	Lower	24	48.79	10.60	- 6.22**	50	0.44
	Higher	28	69.04	12.54			
Speaking	Lower	19	52.21	15.65	- 6.75**	45	0.50
	Higher	28	78.71	11.30			
Writing	Lower	27	52	10.77	- 6.65**	44	0.50
	Higher	19	71.84	8.67			
Overall	Lower	30	54.03	10.09	- 6.89**	58	0.45
	Higher	30	72.80	10.99			

**significant at $p < 0.001$

Table 3 shows that PTE Academic was able to significantly differentiate between the lower and higher proficiency groups as formed by participants' IELTS Academic scores, overall and in each of the four communication skills.

RQ3: *Is there any relationship between participants' PTE Academic scores and their academic achievement as measured by their first semester grade point average (GPA) and as compared to a native-speaker norm group?*

- a. *Is there any significant relationship between participants' PTE Academic scores and their first semester GPA in terms of their overall score and the four communication skill scores?*

Table 4 presents descriptive statistics for the participants' PTE scores and their academic GPA as well as the norm group's GPA.

Table 4. Descriptive statistics for participants' PTE Academic scores, their GPA, and the norm group's GPA

	PTE Listening score	PTE Reading score	PTE Speaking score	PTE Writing score	PTE total score	Participa nts' academic GPA	Native speakers' GPA
Mean	64.65	60.33	66.91	60.50	63.41	70.37	66.57
SD	14.66	14.72	17.10	13.27	14.10	8.26	6.12
Min.	37.00	30.00	10.00	33.00	36.00	46.00	54.60
Max.	90.00	84.00	90.00	88.00	90.00	90.00	80.88

As was also indicated in Table 3, Table 4 shows that participants' highest performance was in the speaking skill (mean = 66.91) followed by their performance in the listening skill (mean = 64.65). Accordingly, we can infer that

this cohort of participants performed better in their PTE Academic oral skills of speaking and listening compared to their writing and reading skills. However, the lowest dispersion was observed in participants' writing performance ($SD = 13.27$), which implies that they were more homogenous in their writing skill.

Table 4 also shows that participants of the study (international students) outperformed native-speaker students as relates to their academic achievement with their mean GPA being 70.37 compared to the native speakers' mean of 66.57. Results of a paired t-test showed that this difference between the two groups' GPA was significant ($t = 2.88$, $df = 59$, $p < 0.01$). Figures 1 and 2 present the distribution of participants' and native speakers' GPA respectively.

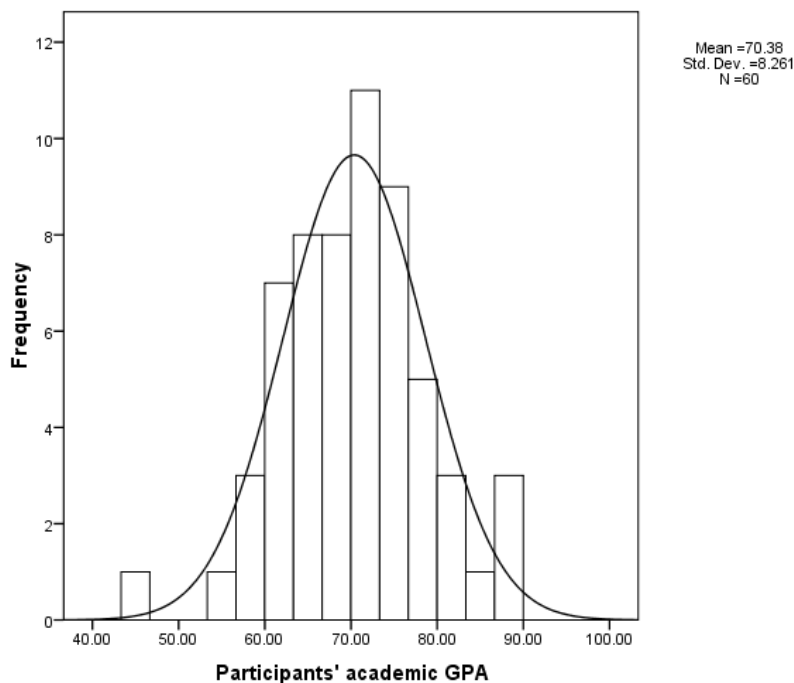


Figure 1. Distribution of participants' (international students) GPA

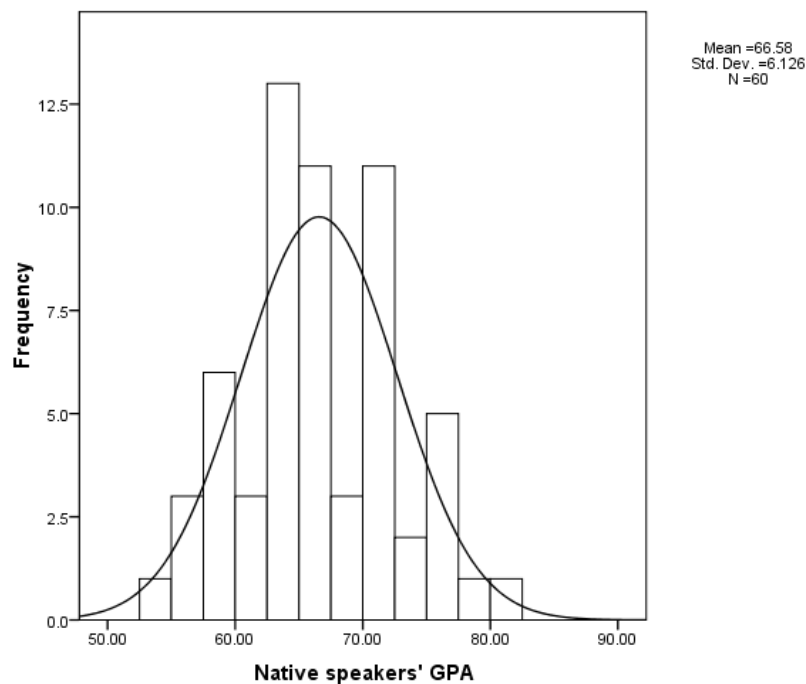


Figure 2. Distribution of native-speakers' (norm-group) GPA

In order to determine if there is any significant relationship between the participants' PTE Academic scores and their academic GPA Pearson correlation was run. Results of the Pearson correlations are presented in Table 5.

Table 5. Relationship between participants' GPA and their PTE Academic scores

	Listening	Reading	Speaking	Writing	Overall PTE score
	0.286*				
		0.350**			
GPA			0.291*		
				0.292*	
					0.344**
r^2	0.082	0.122	0.085	0.085	0.118

*Correlation is significant at 0.05 level (2-tailed)

**Correlation is significant at 0.01 level (2-tailed)

Table 5 shows that the participants' performance on PTE, overall and in each of the four communication skills, had a significant relationship with their academic achievement as measured by their GPA. The highest relationship ($r = 0.350$) was observed between participants' reading scores and their GPA followed by overall score ($r = 0.344$), writing ($r = 0.292$), speaking ($r = 0.291$), and listening ($r = 0.286$).

- b. *What percentage of participants with PTE Academic scores in the A2, B1, B2, and C1 and higher of the CEFR levels reach academic norm as compared to the norm group's GPA?*

In order to answer this research question, the participants were categorised into four CEFR categories using the Pearson PTE-CEFR alignment guidelines (Pearson, 2012). Then, both the international participants' and the norm group's GPA was used to determine the percentage of the participants who reached or exceeded the norm group's GPA. Table 6 presents the number of participants in each CEFR level based on their PTE Academic score and the frequency and percentage of those who reached or exceeded the norm group's GPA.

Table 6. Categorisation of participants' overall PTE Academic scores according to CEFR levels

CEFR Level	PTE Academic score range	Frequency and percentage	Number and percentage of the participants reached or exceeded the norm group's GPA
A2	30–42	4 (6.7%)	2 (50%)
B1	43–58	17 (28.3%)	10 (58.82%)
B2	59–75	29 (48.3%)	19 (65.51%)
C1+	≥76	10 (16.7%)	10 (100%)

Table 9 shows that the largest group of participants (48.3%) were at Level B2 followed by Level B1 (28.3%), C1+ (16.7%), and A2 (6.7%). The last column shows the number of the participants' who reached or exceeded the norm group's academic performance. The percentage increase of participants who reach or exceed the norm group's GPA is directly proportional to their overall language proficiency as represented by their PTE Academic scores. Interestingly, the observed pattern shows that as participants' proficiency goes up from A2 to C1+ they reach or exceed the norm group's academic performance so that all ten participants in the C1+ category outperformed native speakers academically. Table 7 presents details of the participants' and norm-group's GPAs.

Table 7. Participants' and norm-group GPA and percentage of participants reached or exceeded the norm-group's GPA

		Participants' GPA	Norm-group GPA	Percentage of participants reached or exceeded norm-group's GPA
CEFR Levels	A2	63.32	66.68	50%
		73.93	63.36	
		65.00	57.93	
		46.00	59.00	
	B1	71.00	56.25	58.82%
		71.25	58.48	
		58.25	65.50	
		87.25	64.96	
		63.00	80.88	
		61.25	58.21	
		58.50	71.94	
		68.00	69.10	
		59.66	70.62	
		55.33	70.62	
		73.14	55.74	
		80.25	74.39	
		74.62	63.36	
		72.00	64.28	
		69.75	73.54	
		76.00	56.96	
		79.00	64.86	
	B2	65.75	63.98	65.51%
		62.66	60.97	
		71.50	66.03	
		81.75	61.94	
		65.00	70.00	
		71.50	71.50	
		70.25	66.46	
		64.00	65.45	
		71.00	65.00	
		72.50	79.63	
		73.66	70.62	
		72.50	70.62	
		62.00	63.25	
		69.00	66.84	
		78.00	63.06	
		60.25	64.06	
		65.00	77.25	
		84.50	66.64	
		65.75	61.95	
		66.25	75.27	
		76.00	68.00	
		64.33	76.00	
		69.66	65.96	
		76.75	68.42	

	74.30	66.68	
	66.91	64.86	
	60.66	71.62	
	73.50	64.86	
	66.91	65.82	
C1+	72.25	62.93	100%
	78.50	58.92	
	87.71	58.20	
	75.00	71.69	
	75.33	70.62	
	82.00	71.03	
	90.00	77.42	
	68.66	54.60	
	77.33	75.24	
	67.66	64.49	

Further Analysis

Since the same data were available for the IELTS, it was thought it would be interesting to check the relationship between the participants' IELTS scores and their academic performance as well. Table 8 presents the correlations between the participants' GPA and their IELTS Academic, and for the sake of comparison, PTE Academic scores are included as well.

Table 8. Relationship between participants' ($n = 60$) academic performance (GPA) and their PTE Academic and IELTS Academic scores

		Academic performance (GPA)		
		r	r^2	Common variance
PTE Academic	Listening	0.286*	0.082	8.2%
	Reading	0.350**	0.122	12.2%
	Speaking	0.291*	0.085	8.5%
	Writing	0.292*	0.085	8.5%
	Overall	0.344**	0.118	11.8%
IELTS Academic	Listening	0.297*	0.088	8.8%
	Reading	0.340**	0.115	11.5%
	Speaking	0.292*	0.085	8.5%
	Writing	0.362**	0.131	13.1%
	Overall	0.390**	0.152	15.2%

*Correlation is significant at 0.05 level (2-tailed)

**Correlation is significant at 0.01 level (2-tailed)

Both PTE Academic and IELTS Academic tests were functioning almost identically in predicting participants' academic performance as can be observed in Table 8. Both tests indicated significant correlations between participants' GPA and their overall score and the four communication skills scores.

Discussion and Implications

Results of the data analyses provided evidence for both concurrent and predictive validity of PTE Academic. Two sources of evidence supported concurrent validity between PTE Academic and IELTS Academic. The first was the results of correlation between participants' overall scores and the four communication skills of the two tests as presented in Table 2. Significant high correlation was found between the total scores ($r = 0.825, p < 0.001$), and the speaking skill ($r = 0.723, p < 0.001$) of the two tests. These correlations, especially the one related to total scores, provide evidence for the criterion-related validity (concurrent validity) of the PTE Academic with IELTS Academic being the criterion. In terms of common variance between the two tests, the correlation coefficient leads to 68% of common variance between the two tests. An interesting observation from the correlation coefficients is that the correlations for the productive skills (speaking and writing) are higher than the correlations for the receptive skills (reading and listening). This implies that the speaking and writing tasks in the two tests have more common variance compared to the reading and listening tasks.

Another source of evidence, which can provide further support for the concurrent validity of PTE Academic is the results of the independent test presented in Table 3. These results indicate that PTE Academic was able to differentiate between the two proficiency groups (high and low), overall and in the four communication skills, as formed by participants' scores on IELTS Academic. Table 5 also shows medium effect sizes ($\eta^2 = 0.50$) were found for the two productive skills, and small, though close to medium, effect sizes were found for the two receptive and total scores. This implies that those test-takers who achieve higher scores on IELTS Academic will achieve higher scores on PTE Academic, overall and in the four communication skills, and vice versa.

In addition to the evidence found for the concurrent validity of the PTE Academic, further evidence was sought for its predictive validity, i.e., correlations between test-takers' performance on PTE Academic and their academic performance as measured by their GPA. As Tables 5 and 8 show, there were significant relationships between participants' PTE scores, overall and in terms of the four communication skills, and their academic performance as measured by their GPA. The highest relationship ($r = 0.350$) was observed between participants' reading scores and their GPA followed by their overall score ($r = 0.344$), writing ($r = 0.292$), speaking ($r = 0.291$), and listening ($r = 0.286$). In other words, the highest common variance (12.2%) was shown to be between participants' PTE Academic reading score and their GPA which implies that 12.2% of the participants' variance in their academic performance could be

accounted for by their PTE Academic reading scores. Put simply, of the many factors accounting for academic performance, participants' reading ability as represented by their PTE Academic scores is able to predict 12.2% of participants' academic performance. This common variance was found to be 11.8%, 8.5%, 8.5%, and 8.2% for the: overall, writing, speaking, and listening skills, respectively. The implication is that of the four communication skills, reading was able to account for the highest variance (12.2%) in participants' academic performance.

The relatively low but significant relationships among participants' GPA and their overall test scores and the four language skills, as presented in Tables 5 and 8, are in line with findings from similar studies (see, for example, Feast, 2000; Kerstjens & Nery, 2000; To, 2000) on the predictability of IELTS and students' academic performance. These findings demonstrate the predictability of students' academic performance by their English language proficiency test scores within the restricted observed common variances. Although the relatively low correlation coefficients may be a matter of concern for test developers, it should not be surprising given language proficiency is only one of the many factors which could predict academic performance. A myriad of other constructs such as participants' subject matter knowledge, perseverance, study skills, adaptability to the host culture and context, and many other variables (Alderson, Clapham & Wall, 1995) are certainly determinant in participants' academic achievement. Alderson, Clapham and Wall (1995) also state that in predictive validity studies a high correlation should not be expected. They mention that a correlation coefficient of 0.30 would be satisfactory for many predictive validity studies.

Results of further analysis, as presented in Table 8, show that both PTE Academic and IELTS Academic showed almost the same pattern of correlation between participants' test performance and their academic achievement. All correlation coefficients are either higher or around 0.30 providing satisfactory results for the predictive validity of the two language proficiency tests.

Notwithstanding the above observations, test developers may be concerned with improving their tests so that even higher correlations could be obtained between the students' language ability as gauged by language proficiency tests and their academic performance. This is particularly important because the four communication skills (listening, reading, speaking, and writing), along with sub-skills such as comprehension, inferencing, and reasoning these language skills bring about in students, has drawn more attention in present day tertiary education institutions than they did in the past. Movements such as Good Practice Principles (GPP) for English language proficiency for international

students in Australian universities (AALL, 2008) attest to this. Accordingly, two suggestions could be in order here. The first is content-related revisions including test tasks to approach more of the type of academic assignments students do in their university studies. Future revisions of the PTE Academic test may consider how to optimise the design of the test and to include communication tasks that could emulate the type of language-related activities in academia. Such revisions could improve construct under-representation variance of proficiency tests and improve the common variance between the language proficiency tests and students' academic performance. The second suggestion relates to test method facets and particularly test administration issues that could minimise construct-irrelevant factors (Messick, 1989). For example, participants raised some concerns about test administration conditions, however, since the issues raised by the participants are beyond the scope of the current study, the complaints were not collected and analysed in a systematic way to provide valid data and inferences. Nevertheless, the anecdotal narratives are a cause for further concern. It can lead to more systematic research on test-takers' reflections about the test and its administration results of which could help to improve test facets, and accordingly minimise construct-irrelevant variance in test-takers' performance.

Finally, and on the ground that language proficiency is indeed only one of the several predictors of academic achievement, though an important one, it was thought interesting to compare participants' academic performance to a norm group for whom English language proficiency was not an issue. The pattern observed from this comparison, as can be seen in Tables 6 and 7, was that as participants' PTE Academic scores increase from level A2 to level C1+, their GPA reaches or exceeds the norm group's GPA in so far as all the participants' with PTE scores within the C1+ category outperformed the norm group academically as represented by their GPAs. This may have a clear implication for decision-makers in the admissions offices of the universities that students with higher language proficiency will be more successful in their academic subjects. This may indeed impose a trade-off between setting English language cut-scores for admission purposes and the number of students' admitted in universities, but it should certainly concern decision-makers and how they may manage to address this issue. One possibility would be that if students with lower language proficiency levels are admitted, supportive language programs be provided by universities so that international students can improve their language proficiency before they start their mainstream program.

Limitations and suggestions for further research

This study had some limitations which could be used to make suggestions for further research. First, the sample of the study was limited both in terms of quantity and geography which limits the level of generalisation of the findings. Further research with larger samples and recruitment from more diverse populations will certainly provide more evidence for the type of inferences made in this study. The second limitation of this study is its being purely quantitatively-oriented which will result in restricted conclusions. Further qualitative or mixed-methods research projects in which both quantitative and qualitative data and analysis are used should provide more comprehensive conclusions. Closely related to this point, was the fact that participants' views and perspectives on the test overall, its content and structure, and its administration issues were not collected. Future studies could be designed to collect such data and include the analysis in the interpretations of the findings.

Conclusion

Notwithstanding the limitations discussed above, this study provides useful results that contribute to our understanding of a newly developed language proficiency test. The study reported in this paper was conducted as *a posteriori* research intended to seek validity evidence for PTE Academic by focusing on criterion-related validity. The findings of the study complement other studies reviewed in the literature review section and particularly that of Zheng and De Jong's (2011) study, which was an *a priori* study conducted during PTE test development and which sought validity evidence using field test data. In brief, results of the data analyses of the current study showed that:

- There is significant high and moderate correlations between test-takers' overall PTE Academic and their overall IELTS Academic scores ($r = 0.824$), and the four communication skills of: listening ($r = 0.661$), reading ($r = 0.677$), speaking ($r = 0.723$), and writing ($r = 0.686$).
- PTE Academic significantly differentiated between higher and lower proficiency groups as formed by participants' IELTS Academic scores both overall and in all of the four communication skills.
- Significant correlations were found between participants' overall PTE Academic and the four communication skills of listening, reading, speaking, and writing, and their academic performance as measured by their GPA. The highest significant correlation was observed between participants' PTE Academic reading score and their GPA ($r = 0.350$) showing the importance of reading in academic studies. Almost the same

pattern was observed between participants' IELTS Academic scores and their GPA.

- 50% of participants in the A2 level of CERF, 58.82% in B1, 65.51% in B2, and 100% in C1+ categories of CEFR were able to reach or exceed the norm group's (native speakers) GPA; here, all the participants in C1+ group exceeded the norm group's academic achievement.

This last finding has implications for decision-makers in admissions offices of Australian universities given admissions directors would prefer admitting those students who will be academically successful in their studies. The language entry level criterion appears to be one of the predictors for students' future academic performance. The obvious pattern observed from the data is that the percentage of the participants who reach or exceed the norm group's academic performance increases as their PTE scores increase from level A2 to C1+. The relatively low correlations between participants' PTE scores and their GPA, although around the expected index of 0.30 (see Alderson, Clapham, & Wall, 1995), should not affect this observation given the fact that language proficiency is only one of the predictors of academic performance. It should be borne in mind that academic performance also requires academic ability, subject knowledge, perseverance, study skills, adaptability to the host culture and context, and many other variables (Alderson, Clapham, & Wall, 1995) besides language ability.

Acknowledgements

I would like to thank the three research assistants (Olga Kozar, Lily Wang, and Imogen Stewart), but particularly Olga Kozar who helped in the data collection for this project. I also thank the two anonymous reviewers who provided very useful comments on an earlier version of this paper. My special thanks are also due to Macquarie University staff in the Registrar's office who provided me with the academic records. Funding for this research was provided by Pearson Education.

References

- Alderson, J. C. (2009). Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, 26 (4), 621–631.
- Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- AALL (2008). Good Practice Principles for English language proficiency. Retrieved 14 June, 2012, from: <http://aall.org.au/sites/default/files/FinalEnglishLanguageStandardsMay2012.pdf>.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, 2(1), 1–34.
- Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice* (2nd ed.). Oxford: Oxford University Press.
- Chan, S. H. C. (2011). Demonstrating cognitive validity and face validity of PTE Academic writing items Summarize written text and Write essay. Research Note. Retrieved 15 December, 2011, from: http://pearsonpte.com/research/Documents/RN_DemonstratingCognitiveAndFaceValidityOfPTEAcademicWritingItems_2011.pdf.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Educational Testing Service. (2005). *TOEFL® Internet-based test: Score comparison tables*. Princeton: Educational Testing Service.
- Ingram, D. & Bayliss, A. (2007). IELTS as a predictor of academic language performance (Part I). *IELTS Research Reports* (Vol. 7) (pp. 1-68). British Council/IELTS Australia.
- Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, 3(4), 70-85.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Kerstjens, M. & Nery, C. (2000). Predictive validity in IELTS Test: A study of the relationship between IELTS scores and students' subsequent academic performance. *English Language Testing System Research Reports* (Vol. 3) (pp. 85-108). Canberra: IELTS Australia.
- Lu, L. F. (2012). *A validation framework for automated essay scoring systems*. PhD Thesis. Faculty of Education, University of Wollongong. Retrieved 15 March, 2013, from: <http://ro.uow.edu.au/theses/3727>.
- Malone, M. E. (2010). Canadian Academic English Language (CAEL) Assessment. *Language Testing*, 27(4) 631–636.

- Mader, J. & Urkun, Z. (2010). Putting the CEFR to good use. Selected articles by the presenters of the IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG) and EALTA Conference in Barcelona, Spain 29-30 October, 2010. Retrieved 13 November, 2011, from <http://rcel.enl.uoa.gr/kpg/texts/IATEFL%20TEA%20SIG-EALTA%20Barcelona%20Conference%20Proceedings.pdf>.
- Messick, S. (1989). Validity. In R. Linn, (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.
- Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson Test of English Academic. *Educational Research and Evaluation*, 18, 211-229.
- Pearson (2008). Investigating the lexical validity of PTE Academic. Retrieved 12 June, 2011, from <http://pearsonpte.com/research/Pages/ResearchSummaries.aspx>.
- Pearson (2010). Aligning PTE Academic test scores to the Common European Framework of Reference for languages. Research Note. Retrieved 20 March, 2011, from http://pearsonpte.com/research/Documents/Aligning_PTEA_Scores_CEF.pdf.
- Pearson (2012). *PTE Academic Score Guide*. Retrieved 12 August, 2012, from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf.
- To, T. T. H. (2000). *Investigating the predictive validity of the International English Language Testing System (IELTS) Test among Vietnamese students*. Unpublished Ph.D. thesis. The University of Melbourne.
- Uysal, H.H. (2010). A critical review of the IELTS writing test. *ELT Journal*, 64(3), 314-320.
- Wang, H., Choi, I., Schmidgall, J. & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing*, 29(4) 603-619.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Zhang, Y. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET-SET solution. *Language Assessment Quarterly*, 6, 298-314.
- Zheng, Y. & De Jong, J. (2011). Research note: Establishing construct and concurrent validity of Pearson Test of English Academic (PTE Academic). Retrieved 12 June, 2011, from <http://www.pearsonpte.com/research/Pages/ResearchSummaries.aspx>

Appendix 1: Participants' nationality

Nationality	Frequency	Percent
Bangladesh	2	3.3
Brazilian	1	1.7
Chinese	21	35
Filipino	2	3.3
French	1	1.7
German	1	1.7
Hindi	2	3.3
Indonesian	2	3.3
Iranian	2	3.3
Japanese	2	3.3
Korean	4	6.7
Malaysian	3	5
Mexican	3	5.0
Nepali	2	3.3
Pakistani	1	1.7
Peruvian	1	1.7
Russian	1	1.7
Turkish	1	1.7
Vietnamese	8	13.3
Total	60	100.0

Appendix 2: Participants' field of study in each of the three academic disciplines

	Discipline		
	Sciences	Business & Economics	Art & Humanities
Field of study	Biology (1)	Applied finance (5)	Law (2)
	Biotechnology (1)	Business	Media &
	Information technology (3)	administration (4)	International Communication
	Mathematics (1)	Commerce (4)	(6)
	Medical sciences (3)	Economics (2)	Translation &
	Sustainable development (1)	International trade (6)	Interpreting (2)
		Professional accounting (19)	
Total	10	40	10