Research Papers in Language Teaching and Learning
Vol. 12, No. 1, January 2022, 89-110
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Investigating the Parameters Accounting for Language Proficiency: the case of an English language test used for Academic purposes

Dina TSAGARI & Theodosia DEMETRIOU

In language testing it is essential to understand the validity of test scores (Kane 2013 & 2016). However, the focus on whether test items/tasks lead to the target language use expected (Bachman and Palmer, 1996) has been somewhat limited especially in validation studies undertaken by examination boards. The present study explores the (content) validity, e.g. how different linguistic parameters account for language proficiency in high-stakes international English examination for Academic Purposes (the Pearson Test of English Academic - PTE Academic). Various measures of proficiency were taken into account for Writing and Speaking. Results showed that vocabulary mainly accounts for language proficiency and can be used as a predictor variable for the Writing and the Overall Scores in the test. Fluency could also predict some of the variability in the Speaking Scores. The paper contributes to ongoing research on how various language measures can discriminate between levels of proficiency and proposes a statistical model (regression analysis) that can predict speaking and writing scores. Our research intends to provide feedback to test developers or other stakeholders regarding the PTE Academic and offers research and methodological recommendations for the study of content validity of high-stakes tests in other contexts.

Key words: language proficiency; writing; speaking; vocabulary; fluency; regression analysis

---

1. Introduction

In the field of language testing, it is good practice for examination boards to conduct validation studies when making claims about the validity or reliability of their language tests (Weir, 2005). Kane (2013, 2016) suggests that any claims by exam organisations should be supported by evidence to the validity and reliability exam scores while Weir (2005, p. 16) proposes that:

> "Test validation is the process of generating evidence to support the well- foundedness of inferences concerning trait from test scores, i.e., essentially, testing should be concerned with evidence-based validity."

Validation research in the field of language testing has tried to identify variables that affect language proficiency (e.g. Elder & O'Loughlin, 2003; Green, 2009; Hughes Wilhelm, 1997). Motivated by work on the features of spoken and written language (Demetriou, 2016; Banerjee et al., 2007; Barkaoui, 2013; Bosker, 2014; Liontou and Tsagari, 2016; Mayor et al., 2007; Read & Nation 2002; Riazi & Knox, 2013; Seedhouse et al., 2014), the current research study set out to investigate and identify specific speaking and writing features that account for language proficiency in the Global Scale of English (henceforth GSE) used by Pearson Test of English Academic (PTE Academic). Our research intends to use the validation framework proposed by Kane (2015) on validity as a property of score interpretation and provide evidence for writing and speaking test score interpretation in the PTE-Academic.

The study aims at contributing to previous work on how measures of lexical diversity can discriminate between levels (e.g. CEFR levels, see Treffers-Daller, Parslow & Williams, 2016) and offers a statistical model (e.g. regression analysis, using the most important variables/features) that can predict PTE Academic Speaking and Writing Scores. The present study is not an investigation of the relationship between the automated scoring systems employed by Pearson and human raters, but rather an exploration of what is actually captured by the scores allocated. In this study, we are trying to explain which of the variables that are used for the automated scoring are the most important in terms of scoring (in other words, what accounts more for getting higher scores).

The study builds on two main themes: i) the importance of vocabulary in distinguishing between proficiency levels, an area highlighted by various researchers (Crossley, Salsbury, McNamara and Jarvis, 2011a; 2011b; Iwashita et al. 2008; Milton, 2010), and ii) the investigation of other variables accounting for proficiency (Barkaoui, 2013; Griffiths, 1992; Pimsleur et al., 1977; Rimmer, 2006; Tauroza & Allison, 1990; Vanderplank, 1993;). These will be presented in detail in the following section.

2. Importance of vocabulary in distinguishing between language proficiency levels

The importance of vocabulary in distinguishing between proficiency levels (Crossley, Salsbury, McNamara & Jarvis, 2011a; 2011b; Iwashita et al., 2008; Masrai & Milton, 2018; Milton, 2013) and the significant relationship between vocabulary richness and ratings has been highlighted in the literature (Engber, 1995; Lee et al., 2009; Malvern & Richards, 2002; Milton, 2009; Morris & Cobb, 2004; Yu, 2009). Daller and Phelan (2007) also suggest that the use of infrequent words could be an indicator of language proficiency. In addition, the literature also showed that the more words (tokens) produced by a learner, the higher the level they achieved (Morris & Cobb, 2004). Iwashita et al. (2008) found out that the features of vocabulary and fluency as individual detailed features of spoken language produced by test takers have the strongest correlation with levels of speaking performance (also in Crossley et al., 2011a; McNamara, Crossley, & McCarthy, 2010). Adam's (1980) study also showed that vocabulary and grammar were the main components that identified different levels of proficiency.

Furthermore, Hawkey and Barker (2004) concluded that in higher IELTS proficiency levels, essays were longer and employed broader vocabulary. O'Loughlin (2013) analysed PTE Academic writing tasks and found a strong correspondence between the holistic scores of essay responses and academic vocabulary use in terms of tokens and types. Their study showed that frequency

and breadth of academic vocabulary (Academic Word list-AWL tokens and types) were important markers of quality in the essay responses.

In the following studies, almost half of the variation is explained by vocabulary knowledge. For example, research undertaken by Demetriou (2016) on the relationship between vocabulary measures and IELTS Writing Task 2 ratings produced a model that explains nearly 50% of the variance of IELTS Writing Scores and confirms that vocabulary is indeed one of the most important factors with a strong relationship with all other language skills. Schmitt (2010) also reports that findings from previous studies showed that vocabulary accounts for 37-62% of the variance in proficiency scores. Similarly, Crossley et al.'s (2011b) findings revealed that lexical diversity could explain over 45% of the variation in human ratings in general and TOEFL scores in particular.

Treffers-Daller et al. (2016) using PTE Academic essays written by 179 English learners showed that lexical diversity measures can discriminate between CEFR levels. Analysing also learners' scores for each essay, the study showed that the best predictor of CEFR levels that explained 22% of the variance in the scores was the count of words. In addition, basic measures such as the number of different words, Type-Token Ratio (TTR) and Guiraud proved to be better measures than D (Malvern et al. 2004), the Hypergeometric Distribution (HDD - McCarthy & Jarvis, 2007) and Measure of Textual Diversity (MTLD - McCarthy, 2005). Huhta (2014) also found that one of the best predictors for performance on L2 writing tasks were tests of other English skills such as vocabulary.

It is evident from the above empirical research that vocabulary measures are important parameters to consider when investigating the linguistic parameters that account for language proficiency because they can explain a very large percentage of the variance in proficiency ratings/scores.

3. Investigation of other constructs accounting for language proficiency

There is also a large body of literature that investigates other aspects or constructs that account for language proficiency such as fluency, grammar (or grammatical accuracy) and cohesion and coherence. For example, regardless of its problematic definition and its impact on research results (Chambers, 1997; Foster & Skehan, 1996;), it is generally accepted that fluency is one of the descriptors of oral proficiency. Chambers (1997) explains the importance of defining fluency in speech as temporal variables (pauses or lengths of runs between pauses) because it "provides a useful anchorage for a concept which is prone to vagueness and multiple interpretations" (1997, p. 538). As temporal variables are "empirically identifiable and quantifiable", therefore fluency in the current research project will be defined as a temporal variable.

Speech rates or the speed of speech is an additional factor in speaking and understanding English (Griffiths, 1992; Pimsleur et al., 1977; Tauroza & Allison, 1990; Vanderplank, 1993). There are various definitions of speech rate (Chambers, 1997; Levelt, 1989; Riggenbach 1991;). Towell et al. (1996) stress that speech rate alone does not contribute to fluency but other aspects such as frequency of pauses and the length of run (the number of syllables between pauses) are significant factors that need to be taken into account. In addition, Raupach (1987), Towell (1987) and Towell et al. (1996) identified the mean length of runs as the main factor contributing to improvement in fluency.

The importance of fluency in communication has also been highlighted by various other researchers (Foster & Skehan, 1996; Schmitt 1990; Skehan, 1992 & 1996). Most studies on fluency (Ejzenberg, 2000; Freed 1995, 2000; Lennon, 1990; Riggenbach, 1991; Towell et al., 1996) agree that the best predictors of fluency are speech rate (number of syllables articulated per minute) and the mean length of runs (average number of syllables produced in utterances between pauses of 0.25 seconds and above) (Kormos & Denes, 2004, p. 148). In addition, Vanderplank (1993) indicates pacing (the number of stressed words per minute) as another good predictor. Kormos and Denes (2004) also argue that the temporal variable of pace (number of stressed words uttered per minute) is an important predictor of fluency. A more recent study by Révész, Ekiert and Nessa Torgersen (2014) also identified fluency (and lexical diversity) as significant predictors of adequate oral production. The current study also focuses on fluency and investigates whether and how fluency relates to the PTE Academic Scores.

Furthermore, according to Rimmer (2006), grammar also plays an important role in the interpretation of scores. Rimmer's research showed that "grammatical ability correlates highly with overall proficiency and distinguishes between different levels of test-taker performance" (2006, p. 497). Bygate (1999) also stresses that both grammatical accuracy and fluency are important for language proficiency. According to Foster and Skehan (1996), accuracy is defined differently by different researchers (Crookes, 1989; Ellis, 1987; Robinson, Ting, & Urwin, 1995) The current research project study followed the suggestions by Foster and Skehan (1996:304) who recommended that the calculation of error-free clauses has merit as a measure of accuracy. Also the definition of grammatical accuracy used in this study is the proportion of error-free clauses relative to the total number of clauses (Bygate, 1999; Foster & Skehan, 1996; Skehan & Foster, 1997) where an error-free clause is defined as: "a clause in which there is no error in syntax, morphology, or word order" (Foster & Skehan, 1996, p. 310).

Finally, research has highlighted the importance of text features of cohesion and coherence as aspects of proficiency. For example, Banerjee et al. (2007, p. 12) working on cohesive ties, counted all instances of demonstratives (this, that, these, those). Their results showed that "the use of demonstratives seems to tail off at higher levels of language proficiency, suggesting that other cohesive ties come into use" (ibid., p. 61). In the current study, we check if the same applies to PTE Academic Scores. Barkaoui (2013) also found that coherence and cohesion are some of the features that increase in proportion with scores. The researcher operationalised cohesion and coherence using three measures: 'Connectives density', (provides an incidence score for all connectives i.e. causal, additive, temporal and clarification connectives) 'Coreference cohesion' (refers to the phenomenon of when a noun, pronoun or noun phrase refers to another constituent in the text, also in Crossley et al., 2011) and 'Conceptual cohesion' (refers to how semantically or conceptually similar the content of sentences or paragraphs is). These were all calculated using the Coh-Metrix (a software also used in other studies) to provide measurements of various linguistic indices (e.g. Crossley et al., 2011; McNamara et al., 2010; Riazi & Knox, 2013). In the current study, we also use Coh-Matrix and similar operationalisations of cohesion and coherence.

4. Research Questions and Hypotheses

The present study was motivated and informed by the literature reviewed in determining which variables account for language proficiency in the PTE Academic. The study also aims at creating a predictive model consisting of variables that account the most for gaining high PTE Academic Scores. To accomplish its aims, the current research set out to answer the following questions:

1. *Which variables correlate the highest with Scores obtained on PTE Academic?*

We assumed that all variables would correlate with the PTE Academic Scores but since the literature highlights the importance of vocabulary in proficiency ratings, the researchers expected that vocabulary would have more predictive validity than other measures. Therefore, it was expected that vocabulary values would increase in proportion with scores.

2. *Can a statistical model using the variables under investigation explain the variance in the PTE Academic Scores and to what extent?*

Grammar, oral fluency, and written discourse (Coherence and Cohesion) were expected to be features that increase in proportion with scores. A model consisting of variables deriving from all these constructs should be able to predict (to a large extent) the PTE Academic Writing and Speaking Enabling Scores and Overall Scores.

5. Methodology

5.1 Participants and Materials

The participants in the study were 100 learners of English from 11 different countries (See Appendix 1 for countries of origin). 72 of the participants were male and 28 were female. All students took the Pearson Test of English Academic (PTE Academic) and were allocated a particular score for writing and speaking skills. Pearson provided the researchers with Writing essay scripts from PTE Academic with their accompanying scores (Scores for Enabling Skills and Overall Score)[1] and Speaking test recordings with their Scores (scores for Enabling Skills[2] and Overall Score) from students[3].

Overall, the data consisted of 100 essay scripts based on different topics and 200 spoken responses from the same students on two different tasks entitled: 'Describe image'[4]. In such tasks candidates usually describe in detail an image (e.g. chart, graph, picture, table or map) related to an academic theme drawn from the fields of humanities, natural sciences or social sciences.

5.2 Measures

The measures chosen for the operationalisation of the constructs of the grading criteria were decided following suggestions and findings from the studies reviewed. These are presented in Table 1.

[1] These were the test takers' Overall Scores as well as their scores on a number of variables and their Writing and Speaking Score (see PTE Academic Score Guide https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf).

[2] Enabling Scores were provided for: Fluency, Grammar, Pronunciation, Spelling, Textual skills and Vocabulary.

[3] These Speaking and Written Scores are based on all items in PTE Academic which contribute to speaking and written respectively, not only on the essay. Similarly, the Overall Score is based on all items of PTE Academic.

[4] For each exam there are 6 or 7 questions of this type - only two of them were used in this study.

	Variables	Type of Analysis	Method	References
Written Data	Vocabulary	TTR, Guiraud and Guiraud Advanced (a measure of lexical sophistication)	Vocd (CLAN) and Eugene Mollet's programme	Treffers-Daller et al (2016) Demetriou (2016) Liontou & Tsagari (2016)
	Grammatical Accuracy	Error-free clauses, total number of clauses	Manual analysis	Bygate (1999) Foster and Skehan (1996) Skehan and Foster (1997)
	Written Discourse	Cohesion and Coherence	CohMetrix 3.0	Crossley et al. (2011) McNamara, Louwerse, Cai, and Graesser (2005) Graesser, McNamara, Louwerse, and Cai (2004)
		Use of Demonstratives	Manual analysis	Banerjee et al (2007)
		Connectives Density	Incidence score for all connectives i.e. causal, additive, temporal and clarification connectives	Barkaoui (2013)
		Coreference and Cohesion	Argument overlap for adjacent sentences	Barkaoui (2013)
		Conceptual Cohesion	Mean LSA overlap for adjacent sentences and the Mean LSA overlap for adjacent paragraphs	Barkaoui (2013)
Oral Data	Vocabulary	TTR, Guiraud and Guiraud Advanced	Vocd (CLAN) and Eugene Mollet's programme	Treffers-Daller et al (2016) Demetriou (2016)
	Fluency	Rate of Speech	Total number of syllables /total time expressed in seconds multiplied by 60 to give a figure/number expressed in syllables per minute	Kormos and Dénes (2004) Riggenbach (1991)
		Mean Length of Runs	Total number of syllables produced in utterances between pauses of 0.25 seconds and above	Towell et al. (1996)
		Pace	Total number of stressed words per minute	Vanderplank (1993)

Table 1. Overview of measures

5.3 Data Treatment and Analysis

The 100 PTE Academic Writing essays and the 200 test recordings (100 audio files) from the Speaking Part of the exam were extracted from the Pearson database. They were all transcribed and formatted into the CHAT (Codes for Human Analysis of Transcripts, MacWhinney, 2000)

transcription format. The latter is one of the most widely used methods of transcribing oral and written data (MacWhinney, 2000). The data was later analysed using CLAN tools. CLAN (Child Language Analysis) is a program that was designed for the creation and analysis of transcripts in the CHILDES (Child Language Exchange System) database (http://childes.talkbank.org/). It comprises various commands for analysing language including *vocd* (McCarthy and Jarvis, 2007) used also in the present study.

The PTE Academic written essays were subjected to analysis of quantitative measurements for the constructs of Vocabulary, Grammar and Written discourse. The speaking transcripts were subjected to quantitative analysis of Oral Fluency and Vocabulary (see Table 1 for an overview of the measures and methods used).

Vocabulary, Grammar, Written Discourse, and Oral Fluency are 4 of the 6 enabling skills (scoring criteria) used for scoring the PTE Academic. The overall score is based on performance on all test items (tasks in the test consisting of instructions, questions or prompts, answer opportunities and scoring rules). Each test taker does between 70 and 91 items on any given test and there are 20 different item types. For each item, the score given contributes to the overall score. The score range is 10–90 points[5] (See Appendix 2 for the PTE Academic test format and an extract from the PTE Academic Score Guide explaining how the overall score is calculated)[6]. Our aim was to check for correlations between the measurements and the Scores (Enabling skills Scores and Overall Score).

The calculations of the measurements for each construct (based on the transcriptions of the 100 written essays, recordings, and calculations of the various measures) were analysed in SPSS along with the scores received for each essay (data extracted from the Pearson database). Descriptive statistics tested the correlations between the different measures and the test scores. Correlations between measures and scores were calculated and multiple regression analyses produced various statistical models of predictive validity.

6. Results

6.1 Descriptive Statistics

After all the exclusions, 97 Writing, Speaking and Overall Scores were used in the study (out of the original 100). Participants 10, 27 and 95 had to be excluded from the analysis due to insufficient data. The mean score for Writing, Speaking and the Overall Scores are presented in Table 2 which shows the total number of essays and recordings analysed. The mean for the Overall Score was found to be lower compared to Writing and Speaking Scores due to the fact that the calculations of the Overall Score were based on all item scores obtained by every candidate. However, only 2 score items were used in our study, e.g. essay for Writing and description of image task for Speaking.

Descriptive Statistics				
N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**

[5] http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf

[6] The Pearson system uses scoring engines such as the Knowledge analysis Technologies (KAT), Intelligent Essay Assessor (IEA), Reading Maturity Metric (RMM) and Versant Technology.

Writing Score	97	39	88	60.10	12.856
Speaking Score	97	23	90	58.02	16.827
Overall Score	97	33	90	57.64	12.562

Table 2. Descriptive Statistics

Correlations

The pairwise correlations for the Writing Scores and the writing variables can be seen in Table 3.

		Writing Score
Writing Score	Pearson Correlation	1
	Sig. (2-tailed)	
	N	97
Grammatical Accuracy	Pearson Correlation	.552**
	Sig. (2-tailed)	,000
	N	97
Cohesion – Demonstratives	Pearson Correlation	.122
	Sig. (2-tailed)	.235
	N	97
Cohesion – Connectives	Pearson Correlation	-.243*
	Sig. (2-tailed)	.016
	N	97
Cohesion – Coreference	Pearson Correlation	-0.47
	Sig. (2-tailed)	.647
	N	97
Cohesion - Conceptual Sentences	Pearson Correlation	.153
	Sig. (2-tailed)	.135
	N	97
Cohesion - Conceptual Paragraphs	Pearson Correlation	.145
	Sig. (2-tailed)	.156
	N	97
Vocabulary - W - Tokens	Pearson Correlation	.337**
	Sig. (2-tailed)	.001
	N	97
Vocabulary - W - Types	Pearson Correlation	.502**
	Sig. (2-tailed)	.000
	N	97
Vocabulary - W - TTR	Pearson Correlation	.119
	Sig. (2-tailed)	.245
	N	97
Vocabulary - W - Guiraud	Pearson Correlation	.485**
	Sig. (2-tailed)	.000
	N	97
Vocabulary - W - Guiraud Advanced	Pearson Correlation	.426**
	Sig. (2-tailed)	.000
	N	97

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 3. Correlations between the Writing Score and writing variables.

According to Table 3, the measures with the highest correlations are *Grammatical Accuracy* (0.55), *Types* (0.50), *Guiraud* (0.48) and *Guiraud Advanced* (0.42) with statistically significant results and a negative correlation with the *Cohesion-Connectives* (-.24). This result was quite unexpected and will be revisited in the discussion section.

6.2 Correlations between Speaking Scores and Speaking variables

The pairwise correlations for the Speaking Scores and the speaking variables can be seen in Table 4.

		Speaking Score
Speaking Score	Pearson Correlation	1
	Sig. (2-tailed)	
	N	97
Fluency - Rate of Speech	Pearson Correlation	.564**
	Sig. (2-tailed)	.000
	N	97
Fluency - Mean Length of Runs	Pearson Correlation	.060
	Sig. (2-tailed)	.559
	N	97
Fluency – Pace	Pearson Correlation	.442**
	Sig. (2-tailed)	.000
	N	97
Vocabulary - S - Tokens	Pearson Correlation	.501**
	Sig. (2-tailed)	.000
	N	97
Vocabulary - S - Types	Pearson Correlation	.493**
	Sig. (2-tailed)	.000
	N	97
Vocabulary - S – TTR	Pearson Correlation	-.151
	Sig. (2-tailed)	.139
	N	97
Vocabulary - S - Guiraud	Pearson Correlation	.283**
	Sig. (2-tailed)	.005
	N	97
Vocabulary - S - Guiraud Advanced	Pearson Correlation	.136
	Sig. (2-tailed)	.184
	N	97

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 4. Correlations between the Speaking Score and speaking variables.

Table 4 shows statistically significant results of positive correlations between the Speaking Scores and *Rate of Speech* (0.56), *Tokens* (0.50), *Types* (0.49), *Pace* (0.44) and *Guiraud* (0.28) which means that these measures increase when Speaking Scores increase.

To answer our first research question (w*hich variables correlate the highest with scores obtained on PTE Academic*) we also checked the average lexical scores obtained by all candidates and compared writing with speaking. Table 5 presents the descriptive analysis regarding the average lexical scores.

	N	Minimum	Maximum	Mean	Std. Deviation
Vocabulary - W - Tokens	97	118	382	233.58	42.261
Vocabulary - W - Types	97	72	171	123.88	20.485
Vocabulary - W – TTR	97	.33	.68	.5346	.05575
Vocabulary - W - Guiraud	97	5.41	10.31	8.1047	.087575
Vocabulary - W - Guiraud Advanced	97	.24	2.65	1.3834	.54321
Vocabulary - S - Tokens	97	80	257	165.53	35.122
Vocabulary - S - Types	97	41	116	73.92	13.011
Vocabulary - S - TTR	97	.30	.62	.4546	.06608
Vocabulary - S - Guiraud	97	4.35	7.87	5.7655	.68694
Vocabulary - S - Guiraud Advanced	97	.41	2.41	1.198	.38119

Table 5. Descriptive Statistics for Writing and Speaking lexical scores

As can be seen from Table 5 the mean and standard deviation are higher for the Writing lexical scores than the Speaking Scores. In order to check whether this difference was statistically significant, a paired samples t-test was conducted (Table 6).

		Means	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper	t	df	Sig. (2-tailed)
Pair1	Vocabulary-W-Tokens-Vocabulary-S-Tokens	68.052	47.679	4.841	58.442	77.661	14.057	96	.000
Pair2	Vocabulary-W-Types-Vocabulary-S-Types	49.959	19.112	1.941	46.107	53.811	25.745	96	.000
Pair3	Vocabulary-W-TTR-Vocabulary-S-TTR	.07999	.07688	.00781	0.6449	.09549	10.247	96	.000
Pair4	Vocabulary-W-Guiraud-Vocabulary-S-Guiraud	2.33920	.86772	.08810	2.16432	2.51409	26.551	96	.000
Pair5	Vocabulary-W-Guiraud Advanced-Vocabulary-S-Guiraud Advanced	.18530	.60264	.06119	.06384	.30676	3.028	96	.003

Table 6. Paired Samples t-test comparing Writing and Speaking lexical scores

Table 6 shows that the difference between the scores of lexical measures for Writing and Speaking are statistically significant.

7. Regression Analyses and Inference

7.1 Predictive Model for Speaking Scores

After all the correlations were checked, data from the whole population was analysed using multiple regression employing all the previous variables (all fluency and all vocabulary and speaking measures) as predictor variables for the PTE Academic Overall Scores. Since all the measures of Vocabulary and Fluency were used as predictors in the regression model for the Speaking Scores, variance inflation factors were calculated to check the presence of multicollinearity (the phenomenon where two or more predictor variables in regression analysis are highly correlated which means that one can be predicted by the other).

After the first regression, there seemed to be a problem with the variable *Types*. Its Variance Inflation Factor (VIF) was too high (VIF: 420.479). Therefore, another regression was carried out which excluded the variable *Types*. In the second regression, the *TTR VIF* was above 10. To prevent collinearity issues, the value needed to be under 10 (Myers, 1990 in Field 2005). Therefore, the variable *TTR* was excluded (As a rule of thumb, variables higher than 10 were excluded from the model).

A third regression followed without the inclusion of the variable *TTR* making sure that all VIF values were lower than 10. However, another problem arose. The variable *Tokens* had to be excluded from the model because it was not statistically significant (p-value was high, e.g. p=.982). Thus, another regression was performed with the exclusion of *Tokens*. The *Mean Length of Runs* variable was not statistically significant (p=.907) and needed to be excluded. The results of the next regression analysis showed that *Guiraud Advanced* had a high p-value (p=.59). Consequently, it was excluded from the model to improve its validity. The results after the exclusion of *Guiraud Advanced* revealed that *Guiraud* had a high p-value (p=.217) and had to be excluded as well. After these eliminations, only two variables remained in the analysis: *Rate of Speech* and *Pace*. We removed *Pace* because it did not have a significant value either (p=.170). The results of the final regression model (see Table 7) showed that *Rate of Speech* was the variable that could explain 31.9% of the variability in the Speaking Scores (R Square=.319).

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.564*	.319	.311	13.964

* Predictors: (Constant), Fluency – Rate of Speech

Table 7. Final Regression Model (Speaking Score) summary.

Therefore, the fitted regression model for the Speaking Score is as follows:
*PTE Academic Speaking Score: 8.161+ 0.27*Rate of Speech*

7.2 Predictive Model For Writing Scores

The same steps were followed for the creation of the model for predicting the Writing Scores. A regression analysis using backward elimination was carried out using all the writing variables as predictors of Writing Scores.
For the same reasons explained in the previous section (high VIF values), the variables *Types*, *Cohesion-Conceptual Sentences* and *Guiraud* were removed to improve the model. After the last regression, all VIF values were under 10 but some other variables had to excluded because they

were not statistically significant (high p-value), in the following elimination order: *Guiraud Advanced* (p=.095), *Cohesion-Demonstratives* (p=.790), *Cohesion-Coreference* (p=.640) and *Cohesion-Connectives* (p=.118). All the remaining variables were highly significant in predicting the Writing Scores. *Grammatical Accuracy*, *Cohesion-Conceptual Paragraphs*, *Vocabulary Tokens,* and *Vocabulary TTR* were the measures that could explain 50.6% of the variability in the Writing Scores (see Table 8, R Square=.506).

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.712*	.506	.485	9.227

* Predictors: (Constant), Vocabulary – W – TTR, Grammatical Accuracy, Cohesion – Conceptual Paragraphs, Vocabulary – W – Tokens

Table 8. Final Regression Model (Writing Score) summary

Therefore, the fitted regression model for the Writing Score is as follows:
*PTE Academic Writing Score= -43.801+24.493*Grammatical Accuracy + 36.986*Cohesion-Conceptual Paragraphs +0.146*Vocabulary Tokens +90.171*Vocabulary TTR*

7.3 Predictive Model For Overall Scores

Finally, the same procedure was repeated for the creation of the final model for predicting the Overall Scores.

After the last regression analysis, all the VIF values for the remaining variables were lower than 10. However, we excluded variables that were not statistically significant starting with the exclusion of *Cohesion-Demonstratives* (high p-value, e.g. p=.832). After the final regression analysis, all remaining variables were also highly significant in predicting the Overall Scores. *Grammatical Accuracy, Vocabulary –W-Tokens, Vocabulary-W-TTR,* and *Fluency-Rate of Speech* were the variables that could explain 54.5% of the variability in the Overall Scores (see Table 9, R Square=.545).

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.739*	.545	.526	8.651

* Predictors: (Constant), Fluency – Rate of Speech, Vocabulary – W – TTR, Grammatical Accuracy, Vocabulary – W – Tokens

Table 9. Final Regression Model (Overall Score) summary

Therefore, the fitted regression model for the Overall Score was as follows:
*PTE Academic Overall Score= -25.467+24.026*Grammatical Accuracy +0.088*Vocabulary-W-Tokens+ 53.725*Vocabulary-W-TTR+0.110*Fluency Rate of Speech*

8. Summary of results and discussion

In this section each research question will be addressed separately, presenting the results and discussing any implications and relations with the literature reviewed.

The initial assumption with regard to the first research question (*which variables correlate the highest with scores obtained on PTE Academic*) was that all variables would correlate with the

PTE Academic scores. However, since the review of the literature highlighted the importance of vocabulary in proficiency ratings, it was expected that vocabulary would have more predictive validity compared to other measures and, therefore, vocabulary values would increase in proportion with scores.

After the analysis of the Writing Score, the results revealed that the measures with the highest correlations were *Grammatical Accuracy* (0.55), *Types* (0.50), *Guiraud* (0.48) and *Guiraud Advanced* (0.42) with statistically significant results and a negative correlation with the *Cohesion-Connectives* (-0.24). There were three vocabulary variables (*Types, Guiraud* and *Guiraud Advanced*) that correlated with the Writing Scores. This result was not surprising since the expectation was that mostly vocabulary measures would correlate with the Writing Scores. A study by Adam (1980) also showed that grammar and vocabulary were among the main components that distinguished between levels of proficiency. *Grammatical accuracy* seemed to correlate with Scores as well (also in Bygate 1999 and Rimmer 2006) and one of the Cohesion measures, in particular, *Connectives*. This result is also in line with findings from Barkaoui (2013) who also found that measures of Cohesion would correlate with the IELTS Writing Scores.

There are various types of connectives (as explained under Cohesion and Coherence) that are included in the count provided by Coh-Metrix under this index. This includes causal (*as, for, because,* etc.), additive (*and, or so, also, furthermore,* etc.), temporal (*first, finally, meanwhile,* etc.) and clarification (*therefore, as a result,* etc.) connectives. As the literature suggests, the use of these connectives is what makes a text coherent. Therefore, the more of these connectives found in a text, the higher the proficiency level a candidate should be placed at. However, our analysis showed a negative correlation regarding the last variable (*Cohesion-Connectives*) which was quite surprising and unexpected since the negative correlation with the Writing Score means that the more connectives found in an essay the lower the score that essay is assigned to. It could be the case that if there is incorrect use or maybe overuse of these connectives, the essay scores are negatively affected. In other words, the mere existence of these connectives is not enough to account for the score, but their correct use is. Nevertheless, this finding needs further investigation (e.g. the type of connectives) to understand what might have led to this result.

Regarding the Speaking Score, the results revealed positive correlations with the variables *Rate of Speech* (0.56), *Tokens* (0.50), *Types* (0.49), *Pace* (0.44) and *Guiraud* (0.28). Again, as expected here, there are three vocabulary variables (*Tokens, Types,* and *Guiraud*) that correlate with the Scores and two fluency variables. This result once again highlights the importance of vocabulary in proficiency ratings and confirms the hypothesis that vocabulary measures increase in proportion with scores.

The results are also supported by previous studies on the relationship between various variables and proficiency scores. *Tokens* and *Guiraud* were expected to have a correlation with the Overall Scores (Treffers-Daller, 2016; Demetriou, 2016). *Tokens* was also found to be a significant factor that correlated with scores in the studies by Hawkey & Barker (2004), O'Loughlin (2013), and Morris & Cobb, (2014). O'Loughlin (2013), in particular, found a correlation between the variable *Types* and Overall Scores. In addition, it is not surprising that two measures of fluency correlate with the Speaking Scores since the relationship between Fluency and Speaking Scores was also previously highlighted by Iwashita et al. (2008). In particular, it was expected that Speech Rate would be one of the fluency measures that would correlate with the Overall Scores since previous studies (Allison, 1990; Pimsleur et al., 1977; Tauroza & Griffiths, 1992; Vanderplank, 1993) revealed its importance as one of the main factors in speaking English. This was also confirmed in the current study.

As for the second research question (whether there can be a statistical model that can explain the variance in the PTE Academic Scores and to what extent), three predictive models (for Speaking, Writing and Overall Scores respectively) were created using inferential statistics (via SPSS). The three predictive models generated after the analyses conducted are the following:

PTE Academic Speaking Score= 8.161+ 0.27*Rate of Speech
Rate of Speech is the variable that can explain 31.9% of the variability in the Speaking Scores.

PTE Academic Writing Score= -43.801+24.493*Grammatical Accuracy + 36.986*Cohesion-Conceptual Paragraphs +0.146*Vocabulary Tokens +90.171*Vocabulary TTR
Grammatical Accuracy, Cohesion-Conceptual paragraphs, Vocabulary Tokens, and Vocabulary TTR are measures that can explain 50.6% of the variability in the Writing Scores.

PTE Academic Overall Score= -25.467+24.026*Grammatical Accuracy +0.088*Vocabulary-W-Tokens+ 53.725*Vocabulary-W-TTR+0.110*Fluency Rate of Speech
Grammatical Accuracy, Vocabulary –W-Tokens, Vocabulary-W-TTR, and Fluency-Rate of Speech are variables that can explain 54.5% of the variability in the Overall Scores.

The results for the Writing and the Overall Scores were the expected ones. The Writing Score model consists of at least one variable of each of the three constructs under investigation (*Grammatical Accuracy, Vocabulary* and *Cohesion*) and comprised more vocabulary variables (*Tokens* and *TTR*) than the other constructs. The fact that vocabulary seems to account more for the ratings is also supported by previous studies (Treffers-Daller, Parslow and Williams, 2016; Demetriou, 2016) which also showed the significance of these vocabulary measures and their importance for proficiency scores. Furthermore, according to the literature, *TTR* may be affected by text length. However, we did not analyse samples of equal length but opted for the whole essay for each candidate. This was so because previous research by Demetriou (2016) showed that its text dependence flaws make it a good predictor because the better texts are usually longer.
Nevertheless, what was very surprising was the result of the Speaking Scores. There was only one variable in the predictive model for Speaking Scores that accounted for the variability in the score. This was a measure of Fluency (*Rate of speech*). Even though this finding was in line with previous findings from a study by Révész, Ekiert and Nessa Torgersen (2014) who investigated linguistic features for adequate oral production and identified fluency as one of the significant predictors, it was quite surprising that this was the only predictive variable in the present study. One possible explanation for this could be the nature of the task. The nature of language produced in written speech is different from oral speech (Bygate, 2009; 2010; Luoma, 2004;). This could have affected our measurements and results. Furthermore, the 'Describe Image' task is scored on content, fluency, and pronunciation. If the content poorly matches the image, then students will get a low score. Pearson's automated speech recognition system is quite sophisticated and looks also at prosody but this was not investigated in this project.

In addition, time is another important factor. In other words, how quickly one speaks (or how much more one speaks) during a speaking task could be more important than other constructs. Especially on the PTE Academic task used in this study (e.g. 'Describe Image') where the candidates have limited amount of time to speak (25 seconds)[7], it may seem more important to

[7] The Score guide is available at: http://pearsonpte.com/wp-content/uploads/2015/11/PTEA_Score_Guide_05Nov15.pdf

produce as much as one can and as fast as they can. Therefore, the *Rate of speech* could be one of the most significant factors accounting for the Speaking Score.

Another possible explanation could be that for the Speaking measures most calculations were done manually, therefore subjectivity could be an issue here. Most of the fluency measurements for the oral data were extremely hard to transcribe because of the strong accents of most of the participants which made it hard for the researchers to distinguish, particularly the count of stressed syllables. Also, the difficulty that was encountered for calculating *Pace*, for example, could have introduced a measurement error (e.g. uncertainty regarding the relationship between the dependent and independent variables) that might have affected the results. There is also a possibility that the way the transcriptions were prepared (by the researchers rather than the examination board) might have had an impact on the output of the analysis.

Lastly, another possible explanation could be that even though all these measures could work on their own, they cannot be used in combination for the creation of the predictive model. To check this, a new regression analysis was run between the Speaking Score and *Pace*. *Pace* was the last variable that had to be excluded from the model because of its high p-value. Therefore, we used it on its own to check if it would yield different results. The results and model produced can be seen in Tables 10 and 11.

Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	22.109	7.162		3.087	.003		
	Fluency - Pace	.751	.146	.460	5.132	.000	1.000	1.000

* Dependent Variable: Speaking Score

Table 10. Regression Coefficients with variable 'Pace'

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.460*	.212	.204	15.666

* Predictors: (Constant), Fluency – Pace

Table 11. Final Regression Model (Speaking Score) summary

As suspected, even though the p-value is a bit lower than the predictive model of the first Speaking Score, the variable *Pace* has now a significant value and can explain 21% of the variability of the Speaking Score. This means that the faster the candidates speak, the more proficient they are judged. This confirms the hypothesis that even though all these measures are good predictor variables (for the Speaking Scores) when used on their own, they do not work when combined. Therefore, building a statistical model to predict the Speaking Scores may not be as straightforward as once thought. In order to build a statistical model more predictor variables may need to be taken into account and perhaps a more complicated statistical analysis (e.g. checking the residual plots to adjust the model) should be conducted to achieve this.

9. Conclusion

This study focused on the investigation of different features of written and oral speech and their relationship with PTE Academic Scores. Vocabulary seems to be one of the most important variables which accounts for language proficiency and can be used as a predictor variable for the Writing and the Overall Scores. The fact that *Fluency* was the only score that predicted some of the variability of the Speaking Score raised some interesting questions that we addressed but this issue needs to be investigated further. For example, different operationalisations of the fluency construct could be explored, for example adding other aspects of fluency such as repair fluency (Skehan 2003) which was also used in recent studies (Révész, Ekiert, and Nessa Torgersen, 2014). In addition, a larger sample (data based on responses from each candidate on all 6-7 speaking tasks) could be used in the analysis to investigate if this could influence the results. One recommendation for further research would be to obtain and analyse the transcriptions for all the speaking and writing tasks from this dataset (100 candidates) and then compare them with transcriptions and results from the examination board. It may also be interesting to look at participants' L1 and how this might affect the results since the participants' L1 may intervene in how their English as an L2 develops (Murakami, 2016). Finally, it would also be useful to analyse a balanced sample of test-takers responses across different GSE ranges.

In addition, given that the vocabulary lists used in this study for the calculations of *Guiraud Advanced* are not updated, we would like to suggest that future research uses more recent vocabulary lists such as the GSE Vocabulary List (Benigno & DeJong, 2017). The GSE List was created by exploring L1 corpora of spoken and written English and aligning vocabulary to the CEFR and the GSE based on combined criteria of frequency and usefulness. One other venue of research would be the investigation of the use of different words from the List based on candidate responses from different proficiency levels or different L1s. It would also be very interesting to look at the average lexical and other scores obtained by each proficiency group and compare speaking with writing. Lexical diversity scores, for example, are expected to be higher for writing than for speaking. Furthermore, since *TTR* is known to be affected by text length, we would recommend further analysis of samples of equal length to check if the model would remain the same. One last suggestion could be to work on the middle parts of the essays to make sure all essays have an identical number of words (Treffers- Daller et al., 2016).

To conclude, we would like to note that the researchers are aware of the limitations of the study which is not experimental but a construct validity study. Therefore the results of the current study cannot be overgeneralised as the models created here were based on a particular dataset with particular tasks. However, based on cross-sectional data, the study has shown the most significant explanatory composites of L2 speaking and writing skills. Nevertheless, as Ortega and Iberri-Shea (2005) argue, cross-sectional studies cannot be replaced by longitudinal studies for capturing the complex and dynamic processes of L2 development, therefore a longitudinal design can be helpful in understanding such processes (Schoonen et al., 2011). The issues and limitations of this study need to be addressed in further research where different constructs could be added (e.g. spelling or pronunciation) in order to improve the model.

However, the study contributes in identifying the linguistic parameters that account for language proficiency. The findings of the study explain what is actually captured by the allocated scores in the PTE academic and which of the variables that are used for the automated scoring are the most important in terms of scoring, in other words, what accounts more for getting higher scores. According to Pearson, "several proprietary, patented technologies are used to

automatically score test takers' performance on PTE Academic"[8] and it is particularly important for test developers and other stakeholders to be provided with validity evidence for such widely used tests. The intention of our research was to provide feedback to the test developers whether the algorithms used are reasonable, or whether they need to be revisited and if needed amended. Our findings provide important insights into the construct validity of the PTE Academic examination and, in particular, to the construct of writing and speaking. Finally the paper hopes to have provided a model of validation for anyone studying specific correlates of specific skills in high-stakes exam validation.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/ or publication of this article: This work was supported by Pearson Education Ltd, Research Calls 2016.

References

Adams, M. L. (1980). Five Co-occurring Factors in Speaking Proficiency. In J. Firth (Eds.), *Measuring Spoken Proficiency* (pp. 1-6). Washington DC: Georgetown University Press.

Bachman, Lyle. F., and Palmer, Adrian. S. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Banerjee, J., Florencia, F. & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS Band Score levels. *IELTS Research Reports* 7, 241–309. Canberra: IELTS Australia and London: British Council.

Barkaoui, K. (2013). What changes and what doesn't? An examination of changes in the linguistics characteristics of IELTS repeaters' Writing Task 2 scripts. IELTS Research Reports Online Series No 3. Available at https://www.ielts.org/-/media/research-reports/ielts_online_rr_2016-3.ashx

Benigno, V. & J. De Jong (2017). *Developing the GSE Vocabulary*. Technical Report, Pearson. Available at https: //prodengcom.s3. amazonaws.com/GSE-Vocab.pdf

Benigno, V., & De Jong, J. (2017). *Developing the GSE Vocabulary.* Global Scale of English Research Series. Available online https://prodengcom.s3.amazonaws.com/GSE-Vocab.pdf (accessed on 19 February 2021).

Bosker, H. R. (2014). The processing and evaluation of fluency in native and non-native speech. Technical Report. Pearson. Available at http://pearsonpte.com/wp-content/uploads/2014/08/Research-Note_BOSKER-1.pdf

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, *3*(3), 185-214. DOI: 10.1177/136216889900300302

Bygate, M. (2009). Teaching and testing speaking. In Long M. H. and C. J. Doughty *The handbook of language teaching* (pp. 412-440). Oxford, UK: Wiley Blackwell.

Bygate, M. (2010). Speaking. Kaplan, R. B. *The Oxford handbook of applied linguistics* (pp. 63-73). Oxford: Oxford University Press.

Crookes, G. (1989). Planning and Interlanguage Variation. *Studies in Second Language Acquisition, 11*(4), 367-383. DOI: https://doi.org/10.1017/S0272263100008391

[8] https://pearsonpte.com/wp-content/uploads/2021/04/Score-Guide-21.04.21-for-test-takers-1.pdf

Crossley, S.A., Salsbury, T., McNamara, D.S. & Jarvis, S. (2011a). Predicting Lexical Proficiency in Language Learner Texts Using Computational Indices. *Language Testing*, *28*(4), 561-580. https://doi.org/10.1177/0265532210378031

Crossley, S.A., Salsbury, T., McNamara, D.S. & Jarvis, S. (2011b). What Is Lexical Proficiency? Some Answers from Computational Models of Speech Data. *TESOL Quarterly*, *45*(1), 182-193. DOI: 10.5054/tq.2010.244019

Daller, H. & Phelan, D. (2007). What is in a Teacher's Mind? Teacher Ratings of EFL Essays and Different Aspects of Lexical Richness. In Daller, H., Milton J. and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234-244). Cambridge: Cambridge University Press.

Demetriou, T. (2016). *Predicting IELTS Ratings Using Vocabulary Measures*. Unpublished PhD Thesis, University of the West of England, Bristol UK.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In: Riggenbach, H. (Ed.), *Perspectives on fluency* (pp. 287–314). Michigan: The University of Michigan Press.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS Research Reports*, *4*, 207-254. Available at https://www.ielts.org/-/media/research-reports/ielts_rr_volume04_report6.ashx

Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition*, *9*, 12-20. DOI: https://doi.org/10.1017/S0272263100006483

Engber, C.A. (1995). The Relationship of Lexical Proficiency to the Quality of ESL Compositions. *Journal of Second Language Writing*, *4*(2), 139-155. https://doi.org/10.1016/1060-3743(95)90004-7

Field, A. (2005). *Discovering Statistics with SPSS.* London: Sage.

Foster, P. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299-323. https://doi.org/10.1017/S0272263100015047

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second Language Acquisition in a Study Abroad Context* (pp. 123-148). Amsterdam: John Benjamin.

Freed, B.F., (2000). Is fluency, like beauty, the eyes, of the beholder? In Riggenbach, H. (Ed.), *Perspectives on fluency* (pp. 243-265). Michigan: The University of Michigan Press.

Graesser, A., McNamara, D.S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, & Computers*, *36*, 193-202. https://doi.org/10.3758/BF03195564

Green, A. (2009). Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy and Practice*, *14*(1), 75-97. https://doi.org/10.1080/09695940701272880

Griffiths, R. (1992). Speech Rate and Listening Comprehension: Further Evidence of the Relationship. *TESOL Quarterly*, *26*(2), 385-390. https://doi.org/10.2307/3587015

Hawkey, R. & Barker, F. (2004). Developing a Common Scale for the Assessment of Writing. *Assessing Writing*, *9*(2), 122-159. DOI:10.1016/j.asw.2004.06.001

Hughes Wilhelm, K. (1997). Use of an expert system to predict language learning success. *System*, *3*, 317-334. https://doi.org/10.1016/S0346-251X(97)00025-0

Huhta, A. (2014). Diagnosing the development of writing ability. Technical Report. Pearson. Available at http://pearsonpte.com/wp-content/uploads/2014/07/Hutha_A_2014.pdf

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, *29*(1), 24-49. https://doi.org/10.1093/applin/amm017

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, Michael T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1-73. https://doi.org/10.1111/jedm.12000

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.

Lee, Y.W., Gentile, C. & Kantor, R. (2009) Toward Automated Multi-Trait Scoring of Essays: Investigating Links Among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics*, *31*(3), 391–417. DOI: 10.1016/j.system.2004.01.001

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–412. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levelt, W. J. M. (1989) Speaking: From Intention to Articulation. Cambridge, MA: MIT Press.

Liontou, T. and D. Tsagari (2016). 'Integrating Corpus Linguistics and Classroom-based Assessment: Evidence from Young Learners' Written Corpora'. In Tsagari, D. (ed.) *Classroom-based Assessment in L2 Contexts.* Newcastle upon Tyne: Cambridge Scholars Press, pp. 356-382.

Luoma, S. (2004) *Assessing Speaking*. New York: Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, D. & Richards, B. (2002). Investigating Accommodation in Language Proficiency Interviews Using a New Measure of Lexical Diversity. *Language Testing*, *19*(1), 85-104. https://doi.org/10.1191/0265532202lt221oa

Malvern, D., J. B. Richards, N. Chipere, & P. Duran. (2004). *Lexical Richness and Language Development: Quantification and Assessment*. Basingstoke: Palgrave MacMillan.

Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes*, *31*, 44-57.

Mayor, B., Hewings, A., North, S., Swann, J. & Coffin, C. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS academic writing task 2. In L. Taylor and P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 250–313). Cambridge: Cambridge University Press

McCarthy, P. M. & S. Jarvis. (2007). *vocd*: A theoretical and empirical evaluation. *Language Testing*, *24*, 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M. (2005). *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. University of Memphis.

McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). The Linguistic Features of Quality Writing. *Written Communication*, *27*(3-4), 221-246. DOI: 10.1177/0741088309351547

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix version 1.4. Available at http//:cohmetrix.memphis.edu

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 211-232.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *C. Bardel, C. Lindqvist, & B. Laufer (Eds.) L*, *2*, 57-78.

Morris, L. & Cobb, T. (2004). Vocabulary Profiles as Predictors of the Academic Performance of Teaching English as a Second Language Trainees. *System*. *32*(1), 75-87. https://doi.org/10.1016/j.system.2003.05.001

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4), 834-871. https://doi.org/10.1111/lang.12166

Myers, R. (1990). *Classical and Modern Regression with Applications* (2nd Edition). Boston, MA: Duxbury.

O'Loughlin, K. (2013). Research Summary: Investigating lexical validity in the Pearson test of English Academic. Technical Report. Pearson. Available at http://pearsonpte.com/wp-content/uploads/2014/07/OLoughlin_K_2014.pdf

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26-45. DOI: https://doi.org/10.1017/S0267190505000024

Pimsleur, P., C. Hancock & P. Furey. (1977). Speech rate and listening comprehension. In Burt, M., Dulay, H., and Finocchiaro, M. (Eds.). *Viewpoints on English as a Second Language* (pp. 27-34). New York: Regents.

Raupach, M. (1987) Procedural learning in advanced learners of a foreign language. In J. A. Coleman and R. Towell (Eds.) *The Advanced Language Learner* (pp. 123-155). London: CILT.

Read, J. & P. Nation (2002). An Investigation of the Lexical Dimension of the IELTS Speaking Test. *IELTS Research Reports* 6, 207-231.

Révész, A, Ekiert, M & E. Nessa Torgersen (2014). The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance. *Applied Linguistics*, *37*(6), 828-848. https://doi.org/10.1093/applin/amu069

Riazi, A. M. & Knox, J. S. (2013). An investigation of the relations between candidates' first language and the discourse of written performance on the IELTS Academic Writing Test, Task 2. *IELTS Research Reports 2*, 1–89. Canberra: IELTS Australia and London: British Council.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes*, *14*, 423-441. https://doi.org/10.1080/01638539109544795

Robinson, P., Ting, S. C.-C., & Urwin, J. J. (1995). Investigating second language task complexity. *RELC Journal*, *26*, 62-79. DOI: 10.1177/003368829502600204

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 17-46.

Schmitt, N. (2010). Researching Vocabulary: A Vocabulary Research Manual. Basingstoke: Palgrave Macmillan.

Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language learning*, *61*(1), 31-79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Seedhouse P, Harris A, Naeb R, & Üstünel E. (2014). The Relationship between speaking features and band descriptors: a mixed methods study. *IELTS Research Reports Online Series* 2014(2), 1-30. Available at http://eprint.ncl.ac.uk/file_store/production/208149/4BF6D67B-BD63-432D-AA88-2C1E599BBC02.pdf)

Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*, 185-211. https://doi.org/10.1177/136216889700100302

Skehan, P. (1992). *Strategies in second language acquisition*. (Working Papers in English Language Teaching No. 1). London: Thames Valley University.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*, 38-62. http://dx.doi.org/10.1093/applin/17.1.38

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Tauroza, S. & D. Allison. (1990). Speech Rates in British English. *Applied Linguistics, 11*(1), 90-105. https://doi.org/10.1093/applin/11.1.90

Towell, R. (1987) Approaches to the analysis of the oral language development of the advanced learner. In, J. A. Coleman and R. Towell, (Eds.), *The Advanced Language Learner* (pp.157-181). London: CILT.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84-119. https://doi.org/10.1093/applin/17.1.84

Treffers-Daller, J., Parslow, P. & Williams, S. (2016). Back to Basics: How measures of Lexical Diversity can help discriminate between CEFR Levels. *Applied Linguistics*, *37*(4), 1-27. https://doi.org/10.1093/applin/amw009

Vanderplank, R. (1993). Pacing and spacing as predictors of difficulty in speaking and understanding English. *English Language Teaching Journal*, *47*, 117-125. https://doi.org/10.1093/elt/47.2.117

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

West, M., & West, M. P. (Eds.). (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Addison-Wesley Longman Limited.

Wray, A. (2002). *Formulaic Language and the Lexicon.* United States of America, New York: Cambridge University Press.

Yu, G. (2009). Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*, *31*(2), 236-259. https://doi.org/

Appendix 1

Countries of origin and number of participants

Countries	Number of Participants
India	44
Australia	41
United States	4
United Arab Emirates	4
Singapore	2
Nigeria	1
Indonesia	1
Turkey	1
United Kingdom	1
Puerto Rico	1
Jordan	1

Appendix 2

PTE Academic content and scoring system - Test Format

Part 1: Speaking & Writing (77 – 93 minutes)

- Personal introduction
- Read aloud
- Repeat sentence
- Describe image
- Re-tell lecture
- Answer short question
- Summarize written text
- Essay (20 mins)

Part 2: Reading (32 – 41 minutes)

- Multiple choice, choose single answer
- Multiple choice, choose multiple answers
- Re-order paragraphs
- Reading: Fill in the blanks
- Reading & writing: Fill in the blanks

Part 3: Listening (45 – 57 minutes)

- Summarize spoken text
- Multiple choice, choose multiple answer
- Fill in the blanks
- Highlight correct summary
- Multiple choice, choose single answer
- Select missing word

- Highlight incorrect words
- Write from dictation

Scoring system

The overall score is based on performance on all test items (tasks in the test consisting of instructions, questions or prompts, answer opportunities and scoring rules). Each test taker does between 70 and 91 items in any given test and there are 20 different item types. For each item, the score given contributes to the overall score. The score range is 10–90 points.

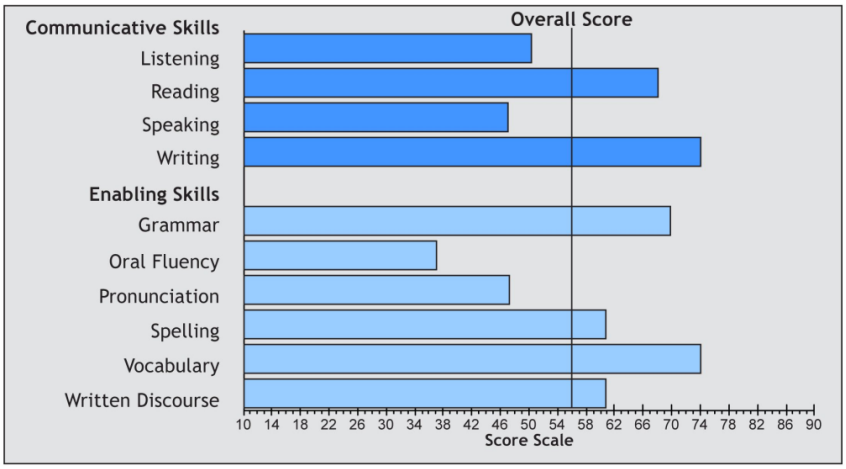
Overall Score: 56

The Overall Score for the PTE Academic is based on the test taker's performance on all items in the test. The scores for Communicative Skills and Enabling Skills are based on the test taker's performance on only those items that pertain to these skills specifically. As many items contribute to more than one Communicative or Enabling Skill, the Overall Score cannot be computed directly from the Communicative Skill scores or from the Enabling Skill scores. The graph below indicates this test taker's Communicative Skills and Enabling Skills relative to his or her Overall Score.

When comparing the Overall Score and the scores for Communicative Skills and Enabling Skills, please be aware that there is some imprecision in all measurement, depending on a variety of factors. For more information on interpreting PTE Academic scores, please refer to *Interpreting the PTE Academic Score Report* which is available at www.pearsonpte.com/pteacademic/scores.

Skills Profile

Communicative Skills	
Listening	51
Reading	69
Speaking	47
Writing	74
Enabling Skills	
Grammar	70
Oral Fluency	37
Pronunciation	47
Spelling	61
Vocabulary	74
Written Discourse	61



Dina Tsagari (dina.tsagari@oslomet.no) is a Professor at the Department of Primary and Secondary Teacher Education, Oslo Metropolitan University, Norway. Her research interests include language testing and assessment, materials design and evaluation, differentiated instruction, multilingualism, distance education and learning difficulties. She is the editor and author of numerous books, journal papers, book chapters, project reports etc. She coordinates research groups, e.g. CBLA SIG – EALTA, EnA OsloMet and is involved in EU-funded and other research projects (e.g. NOHED, KriT, DINGLE, ENRICH, TALE, DysTEFL, PALM, etc)

Theodosia Demetriou (demetriou.th@unic.ac.cy) is a Lecturer at the Department of Education at the University of Nicosia, Cyprus. Her area of expertise is Language Testing and Assessment and her research interests include vocabulary measurement, dimensions of vocabulary knowledge, language testing and assessment and predictive models for IELTS scores. She worked as a Lecturer of English and Linguistics at the University of the West of England, Bristol and as a postdoctoral researcher at the University of Cyprus, (2017). She is currently participating in various research projects in teacher's assessment literacy in Greece and Cyprus, perceptions of senior academics towards research evaluation and multilingualism and identity styles in Cyprus.