ISSN 2515-1703
2021/2

IELTS Partnership Research Papers:
Studies in Test Comparability Series

Investigating the relationship between IELTS Academic and PTE-Academic

Edited by Nick Saville, Barry O'Sullivan and Tony Clark

IELTS™
BRITISH COUNCIL idp Cambridge Assessment English

Investigating the relationship between IELTS Academic and PTE-Academic

This volume of *Studies in Test Comparability Series* contains two studies which offer test score users an opportunity to draw on two analytic approaches when making comparisons between IELTS Academic and PTE-A. These perspectives encourage prospective test score users to move beyond the basic comparison of overall scores to a more nuanced awareness of underlying similarities and differences.

Funding

This research was funded by the British Council and supported by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Publishing details

Published by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia © 2021.

How to cite this volume

To cite this edited volume:
Saville, N., O'Sullivan, B, & Clark. T. (Eds.) (2021). Investigating the relationship between IELTS and PTE-Academic. *IELTS Partnership Research Papers: Studies in Test Comparability Series*, No. 2. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

To cite the first study in this volume:
Yu, G. (2021). IELTS Academic and PTE-Academic: Degrees of Similarity. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series*, No. 2, (pp. 7–41). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

To cite the second study in this volume:
Elliot, M., Blackhurst, A., O'Sullivan, B., Clark, T., Dunlea, J., & Saville, N. (2021). Aligning IELTS and PTE-Academic: A measurement study. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series,* No. 2, (pp. 42–64). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Foreword

The two studies contained in this report offer test score users an opportunity to draw on two analytic approaches when making comparisons between IELTS Academic and PTE-A. These perspectives encourage prospective test score users to move beyond the basic comparison of overall scores to a more nuanced awareness of underlying similarities and differences.

Institutions should consider a range of evidence when setting standards for their specific purposes, as the range of activities sampled by different tests (and the depth in which they do so) differs. As such, the applicability of scores may vary, depending on the range of activities in which applicants will typically be engaged.

Making comparisons between scores on different tests is challenging because tests differ in their design, purpose and format (Taylor, 2004, Lim et al, 2013), and the greater the difference in design, the more problematic the comparison is. Nonetheless, test score users are often interested to know how results on two differing tests may compare.

The two separate reports, each reflecting a different methodology, highlight the need to consider any equivalence estimate from two distinct perspectives:

1. Construct
2. Measurement.

The Construct approach typically entails a detailed evaluation of the way in which the tasks and items contained in the test reflect the target language construct. For test scores to be truly meaningful, we do not simply focus on the language. Instead, we broaden our focus to the underlying cognitive demands of the test tasks (do they reflect those of the real world of language use) while understanding the impact of the social conditions of language use which is particularly relevant for the productive skills, where social parameters such as speaker/interlocutor relationship is always likely to impact on performance.

The Measurement approach compares the scores achieved across the different sections of the test. This allows us to draw comparisons around the measurement relationship between the two, for example, allowing us to answer questions such as how well one test can predict performance on the other.

By combining two studies, we hope to give readers a understanding of the relationship between the two tests under investigation than would be the case if only one approach were taken.

A brief overview of the construct study: *IELTS Academic and PTE-Academic: Degrees of Similarity*

The first study reported here was commissioned by the IELTS Partners, and focuses on a comparison of IELTS Academic and PTE-A. Professor Guoxing Yu uses Kolen and Brennan's (2014) Degrees of Similarity (DES) framework to offer a broad comparison between the two tests. He also applies Weir's (2005) socio-cognitive framework as the basis of a holistic exploration of test task performance parameters. In addition, Yu interviewed individuals who had taken both tests in order to gain additional insight into their experiences and observations.

Yu defines the four test features that form the DES framework as:

- Populations: To what extent are the two tests designed to be used with the same populations?
- Constructs: To what extent do the two tests measure the same constructs?
- Measurement characteristics/conditions: To what extent do the two tests share common measurement characteristics or conditions including, for example, test length, test format, administration conditions, etc.?
- Inferences: To what extent are scores for the two tests used to draw similar types of inferences?

Population

Based on the similarity of the target test-taker populations, Yu suggests that it is feasible to compare the two tests and that we should expect significant overlap across the tests in terms of the construct and measurement characteristics and conditions.

Constructs and Measurement characteristics/conditions

Yu concludes that the speaking tests are very different in terms of how they assess the skill and what aspects of the skill are tested. The lack of publicly available information on how PTE-A estimates overall ability in speaking makes comparison difficult. While a similar situation was reported for the PTE-A writing paper, Yu also finds that the structure of that paper compromised his ability to draw meaningful comparisons. As for the receptive skills, Yu saw little overall difference across the listening papers, though felt that the reading papers were quite different. Here he suggests that the PTE-A reading paper is somewhat less demanding than the IELTS Academic reading paper, though acknowledges that the difference is not considerable.

Inferences

In his conclusions, Yu states that while it is feasible that the inferences drawn from test performance is generally similar for both tests, there are a number of issues that test score users should take into consideration when deciding on which test is suitable for use in their specific context.

A brief overview of the measurement study: *Aligning IELTS and PTE-Academic: A measurement study*

The data used in this study were obtained by Catalyst Research of Perth, Australia, as part of a survey of test-taker experiences with different tests. Score information was obtained from 523 test-takers who had taken both tests within 90 days of each other. The majority had taken IELTS in Australia and represented a range of nationalities/first language backgrounds, including Chinese, Indonesian and Polish, while smaller numbers had taken IELTS in Hong Kong, Pakistan and the UK. Only 115 participants provided their overall scores, so analysis at individual skill level is based on just 408 test-takers. The first analysis undertaken was a simple correlation between performance on the two tests, i.e. how far they agree in their rank-ordering of the test-takers. This is of interest because it points to the extent to which the tests can be regarded as testing the same construct (the range of performances that the test's design seeks to assess and the tasks employed to do this).

The findings from this analysis indicate that the overall equivalences reported here and in a recent report from Pearson (Clesham & Hughes, 2020) are very similar in that both highlight the weakness of the relationship across the two speaking papers.

Additional analysis was undertaken using Equipercentile linking with pre-smoothing, as described in Kolen and Brennan (2004). This approach to smoothing is advantageous in that indices are available for evaluating the goodness of fit and therefore of the linking. The linking was carried out using the RAGE-RGEQUATE software (Zeng, Kolen, Hanson, Cui & Chien, 2004).

Findings highlighted quite significant differences across the two productive skills, and highlighted the need to move away from a solitary focus on the single overall score data as this approach can mask important differences between the tests.

Professor Barry O'Sullivan
British Council

References

Clesham, R., & Hughes, S. R. (2020). *2020 Concordance Report PTE Academic and IELTS Academic*. Accessed 14 January 2020 from: https://pearsonpte.com/wp-content/uploads/2020/12/2020-concordance-Report-for-research-pages.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer Science & Business Media.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice, *International Journal of Testing*.

Taylor, L. (2004). Issues of test comparability, *Research Notes 15*, 2–5.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Palgrave Macmillan.

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2004). RAGE-RGEQUATE [Computer software]. Iowa City: University of Iowa.

Contents

REPORT 1, IELTS Academic and PTE-Academic: Degrees of Similarity 7

Abstract 7

Author biodata 8

1. Introduction 9

2. Overview of the two tests 9

2.1 IELTS: Paper-based and computer-delivered 9

2.2 Pearson Test of English Academic 9

3. Analytic frameworks: A brief introduction 10

3.1 Degrees of similarity 10

3.2 Socio-cognitive framework 11

4. Data and methods of analysis 11

5. Findings 12

5.1 Populations 12

5.2 Constructs and measurement characteristics/conditions 14

5.3 Inferences 32

6. Discussions and conclusion 34

References 38

Appendix 1: Interviews with IELTS and PTE test-takers 41

REPORT 2, Aligning IELTS and PTE-Academic: A Measurement Study 42

Abstract 42

Authors' biodata 43

1. Introduction 46

2. Aligning tests 46

2.1 Quantitative-only studies 47

2.2 Qualitative and quantitative studies 48

3. The current study 49

4. Methodology 49

4.1 Participants 49

5. Analysis 50

6. Results 51

6.1 Scatterplots 51

6.2 Equipercentile graphs 53

6.3 Comparing the current study with Clesham & Hughes (2020) 58

6.4 An alternative alignment table 59

7. Conclusions 60

7.1 Interpreting results across concordance tables 60

7.2 Integrating quantitative and qualitative data: Summarising the results of the current study and Yu (2021) 61

7.3 Limitations 61

References 62

REPORT 1
IELTS Academic and PTE-Academic: Degrees of Similarity

Guoxing Yu

Abstract

Kolen and Brennan (2014) suggested that 'the utility and reasonableness of any linking depends upon the degree to which tests share common features' (p.498) as a starting point for any linking or alignment exercise. They suggested considering at least four features in examining similarity: populations, constructs, measurement characteristics/ conditions, and inferences.

Following Kolen and Brennan's Degrees of Similarity framework and utilising Weir's (2005) socio-cognitive framework, we analysed the official sample questions/tasks of IELTS Academic and PTE-Academic, and various promotional and research publications by and/or on IELTS and PTE. In addition, we conducted semi-structured interviews individually with three candidates who have taken both IELTS and PTE multiple times.

It is evident that the two tests serve the similar populations and purposes and have some commonalities in the underlying constructs of the four language skills. However, the operationalisation of the constructs varied to a large extent. Several assessment methods are unique to PTE; for example, integrated assessment is a prominent feature of several PTE tasks (e.g. summarise written text, summarise spoken text, retell lecture, and describe image), which are also linguistically and cognitively more demanding than other tasks. The difficulty level of IELTS tasks is more balanced across the papers but the difficulty level of the PTE tasks varies to a greater extent within a paper. Some PTE tasks look more authentic, academic-oriented, and demanding, but their difficulty might be cancelled out by easier tasks which assess mainly, if not solely, lexical knowledge and local-level comprehension of the inputs.

The overall cognitive and linguistic demands of the two tests are broadly similar, though there are variations between different papers (Speaking, Writing, Listening, and Reading). Another prominent difference between the two tests is in relation to the transparency of the weightings of different question types and different skills in the calculation of the overall score/band. IELTS provides all the information about its scoring methods and the weighting of each question and task. The biggest challenge in identifying the degrees of similarity between the two tests is caused by the lack of information about PTE on the weightings of different question types and the weightings of different skills in the integrated assessment tasks to calculate the overall score and the six enabling skills scores.

The findings of our textual analyses urge for more fine-tuned equivalence tables which should incorporate not only the overall scores/bands, but also the four language skills separately at different band/score level, and even at a question/task type or a set of similar question/task types, to reflect the big differences in constructs and measurement characteristics between the two tests. In addition, we suggest that any equating exercise should engage with, and collect more, qualitative data from key stakeholders such as test-takers, teachers of test preparation courses, and test score users. The fine-tuned equivalence tables, incorporating both correlational statistics and qualitative data from key stakeholders would facilitate test score users to make more informed inferences about the test results.

Contents: Study 1

1. Introduction **9**

2. Overview of the two tests **9**

2.1 IELTS: Paper-based and computer-delivered 9

2.2 Pearson Test of English Academic 9

3. Analytic frameworks: A brief introduction **10**

3.1 Degrees of similarity 10

3.2 Socio-cognitive framework 11

4. Data and methods of analysis **11**

5. Findings **12**

5.1 Populations 12

5.2 Constructs and measurement characteristics/conditions 14

5.3 Inferences 32

6. Discussions and conclusion **34**

References **38**

Appendix 1: Interviews with IELTS and PTE test-takers **41**

List of tables

Table 1: Overview of IELTS Speaking tasks 15

Table 2: Overview of PTE Speaking tasks 16

Table 3: Summary of linguistic and cognitive processing demands of the Speaking tasks 19

Table 4: Overview of IELTS Writing tasks 19

Table 5: Overview of PTE Writing tasks 20

Table 6: Summary of linguistic and cognitive processing demands of the Writing tasks 22

Table 7: Overview of PTE Listening tasks 24

Table 8: Summary of the linguistic and cognitive demands of the Listening tasks 27

Table 9: Examples of IELTS Reading passages at CEFR level 28

Table 10: Overview of PTE Reading tasks 29

Table 11: Summary of the linguistic and cognitive demands of the Reading tasks 31

Table 12: PTE and IELTS equivalence as reported by Pearson 33

Table 13: Example of IELTS and PTE entry requirements by a competitive UG program 33

Author biodata

Professor Guoxing Yu, University of Bristol, earned his PhD from Bristol in 2005, supervised by Prof. Pauline Rea-Dickins; his dissertation was awarded the Jacqueline Ross TOEFL Dissertation Award by Educational Testing Service (2008). He is an Expert Member of European Association for Language Testing and Assessment; an Executive Editor of Assessment in Education; a member of Editorial Board of *Assessing Writing, Language Assessment Quarterly, Language Testing*, and *Language Testing in Asia*, and the co-editor of two book series: P*edagogical Content Knowledge for English Language Teachers* (Foreign Language Teaching and Research Press, with Peter Gu, Victoria University of Wellington) and *Research and Practice in Language Assessment* (Palgrave, with Anthony Green, University of Bedfordshire). He has published widely in academic journals including *Applied Linguistics, Assessing Writing, Assessment in Education, Language Testing, Language Assessment Quarterly*, and *IELTS Research Reports*.

How to cite this study:

Yu, G. (2021). IELTS Academic and PTE-Academic: Degrees of Similarity. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series*, No. 2, (pp. 7–41). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

1. Introduction

In order to investigate the comparability of IELTS Academic (hereafter IELTS) and Pearson Test of English Academic (hereafter PTE) test results, the Degrees of Similarity (Kolen, 2007; Kolen & Brennan, 2014, pp.498–500) and Weir's (2005) socio-cognitive frameworks were adopted to compare four aspects of the two tests – constructs, inferences, populations, and measurement characteristics/conditions. In addition to the official sample questions/tasks provided by the two tests on their official websites or apps, another two sources of data were analysed: (a) semi-structured interviews with three candidates individually (more than three hours) who have taken both IELTS and PTE multiple times to meet their respective purposes – admission to competitive undergraduate programs and/or application for Australian immigration: and (b) various promotional and research publications by or on IELTS and PTE.

2. Overview of the two tests

2.1 IELTS: Paper-based and computer-delivered

IELTS is an international test of English proficiency assessing all four skills – Listening, Reading, Speaking, and Writing. It has been in operation for 30+ years. The British Council, IDP: IELTS Australia and Cambridge Assessment English jointly own IELTS. There are two types of IELTS test: IELTS Academic and IELTS General Training. The Listening and Speaking papers are the same for both IELTS tests, but the Reading and Writing papers are different. The Listening, Reading, and Writing papers are completed in one sitting, without breaks. The Speaking test is completed separately, either within a week or so before or after the written test. The total test time is 2 hours and 45 minutes, in the sequence of Listening, Reading, and Writing in one sitting, plus the Speaking test in a separate sitting as described above.

IELTS is a primarily paper-based test, but it is also offered in a computer-delivered format. Computer-delivered IELTS is the same as the paper-based IELTS in terms of content, structure, question types, marking, test report form, and test timings. However, the test timing for Listening is slightly different. In the paper-based IELTS, test-takers need to transfer their answers to an answer sheet, while this step is unnecessary in computer-delivered IELTS when test-takers can answer directly on computer. The Speaking test remains face-to-face with a certified IELTS examiner in computer-delivered IELTS.

Test results are reported on a scale of 0–9 for the four skills separately as well as an average score for the whole test, which is also reported as a Common European Framework of Reference for Languages (CEFR) level. Test results are made available within 3–5 days for computer-delivered IELTS, and on the 13th day for paper-based IELTS. The Test Report format remains the same for computer-delivered IELTS, and paper-based IELTS.

2.2 Pearson Test of English Academic

PTE Academic is a computer-based test (Wang et al, 2012). It was launched in 2009. It takes about 3 hours to complete; candidates are given a slightly different number of items/tasks to complete (see more details in Section 5: Findings). It has three parts: Part 1, Speaking and Writing (77–93 minutes); Part 2, Reading (32–40 minutes); Part 3: Listening (45–57 minutes). Parts 1 and 3 contain several question types which are all individually timed. Part 2 (Reading) is timed as a paper. There is an untimed introduction to the test before Part 1 and one optional scheduled break of up to 10 minutes between Part 2 and Part 3. According to PTE official reports (e.g. Pearson 2019b), there are 20 different question types in total in the test (note: the same multiple question type for Reading and Listening is counted as two different question types in this calculation).

All items are machine scored using automated scoring systems. There are two types of scoring: correct or incorrect, and partial credit. Scores are reported on a scale of 10–90: six enabling skills (grammar, oral fluency, pronunciation, spelling, vocabulary, written discourse); four communicative/language skills (reading, writing, listening, speaking); and one overall test score (note: the overall test score is not exactly the average of the scores of the four communicative skills (see Section 5.3: Inferences, and the test scores of the three interviewees).

Test results are normally available within 5 business days. Currently, test results are 'typically available within just 48 hours of taking the test' as PTE states on its official website. Many test-takers receive their test results within the same day, as one of the interviewees did. Test results can be sent to as many institutions as test-takers like, without an additional fee. As with IELTS, PTE test results are used for a range of purposes, including university admission, migration applications, and registration for professional associations.

3. Analytic frameworks: A brief introduction

3.1 Degrees of similarity

Kolen and Brennan (2014) argued that one way to think about linking any two tests is 'in terms of degrees of similarity' in test features (p.498). They also argued that 'the utility and reasonableness of any linking depends upon the degree to which tests share common features' (p.498) as a starting point for any linking or alignment exercise. They suggested considering at least four features in examining similarity: populations, constructs, measurement characteristics/conditions, and inferences.

- Populations: 'To what extent are the two tests designed to be used with the same populations?' In other words, who are the intended users – test-takers and other score users?
- Constructs: 'To what extent do the two tests measure the same constructs?' In other words, 'whether the true scores for the two tests are functionally related'. It is very likely that two tests may share some common constructs, but they also have their unique constructs.
- Measurement characteristics/conditions: 'To what extent do the two tests share common measurement characteristics or conditions including, for example, test length, test format, administration conditions, etc?' (p. 498). Measurement characteristics/conditions are the actual manifestations of test constructs in concrete terms, which often can be understood from test specifications and their operationalisation in test tasks. Measurement characteristics/conditions therefore, in effect, refers to all aspects or facets of a test.
- Inferences: 'To what extent are scores for the two tests used to draw similar types of inferences?' In other words, 'whether the two tests share common measurement goals' in a scale(s) designed to yield similar types of inferences about test results. If two tests differ with respect to their intended inferences, it would mean that 'the two tests were developed and are used for different purposes' (p.499).

3.2 Socio-cognitive framework

Weir's (2005) socio-cognitive framework for language assessment was adopted to investigate in detail the degree of similarity between IELTS and PTE, along with the Degrees of Similarity framework (Kolen, 2007; Kolen & Brennan, 2014, pp.498–500). Various other publications utilising the socio-cognitive model were also used (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011; Taylor & Weir, 2012; Weir, Vidaković & Galaczi, 2013).

The socio-cognitive framework comprises several construct-related components, including test-taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity. The socio-cognitive framework allows for a systematic theoretical consideration of the different aspects of validity but can also offers an important checklist for conducting analysis of test content more practically.

- Test-taker characteristics refers to who the target candidates are, how candidates' physical/physiological, psychological, and experiential characteristics are catered for by a test or task, and whether the test or task is appropriate for the target candidates.
- Cognitive validity concerns the extent to which the cognitive processes required to complete the tasks correspond to the underlying theoretical construct of language ability, as well as the extent to which the same cognitive processes would be involved when completing the same/similar task in the target real-life language use context.
- Context validity refers to the extent to which a test or task is capable of eliciting language samples that are representative of the real-life construct under measurement.
- Scoring validity refers to the extent to which the scores derived from the test are consistent, reliable and can be generalised to real-life language use context. It also concerns the appropriateness of the rating criteria.
- Consequential validity refers to the extent to which test scores are interpreted and used as intended; as well as the potential consequences or impacts of test score use and interpretation on teaching, learning and the society more generally.
- Criterion-related validity refers to the extent to which the relationships of the test with external sources can be established to support the way by which the test scores are derived and used as intended.

4. Data and methods of analysis

Three sources of data were analysed to investigate the degree of similarity between the two tests.

- The handbooks, sample questions/tasks and other test preparation materials provided by the two tests on their official websites or apps. In addition, Cambridge IELTS Volumes 13 and 14 were analysed. Even though there are so many different test preparation materials or apps available these days, we focused on the official materials only. TextInspector was used to analyse the text inputs (mainly reading passages which meet the threshold number of words to run meaningful and reliable textual analysis) to identify their various lexical features and CEFR level more generally.
- Semi-structured individual interviews with three candidates, who have taken both IELTS and PTE multiple times to meet their respective purposes – admission to competitive undergraduate programs and/or application for Australian immigration. Each interview lasted just over one hour and was conducted in their first language.

The semi-structured interview focused on eight series of questions (see Appendix 1) on their reasons for taking both tests, their experience in preparing for and taking the tests, and their views on the similarity and difference between the two tests, their understanding about the overall difficulty level of the two tests, the specific challenges they faced in preparing for and answering different types of tasks, and their suggestions on how PTE and IELTS might learn from each other on test design and delivery.

- Research publications and promotional materials by or on PTE[1] and IELTS[2], which provide further evidence on the degree of similarity and difference between the two tests, from the perspectives of researchers and test providers, respectively.

Textual analysis is conducted on the official sample questions/tasks, utilising the degrees of similarity framework in conjunction with the socio-cognitive framework. The rest of the data – interviews, research articles and promotional materials – are analysed thematically and quoted in this report where appropriate to support the interpretations of the findings from the textual analysis of the official sample questions/tasks.

1. See https://pearsonpte.com/organizations/researchers/external-research-projects/
2. See https://www.ielts.org/teaching-and-research/research-reports
3. For example, https://immi.homeaffairs.gov.au/help-support/meeting-our-requirements/english-language

5. Findings

5.1 Populations

For each test, the population is the group of candidates for whom the test is intended. For this aspect of the analysis, we aim to identify what population each test is designed to assess and what population is being assessed, and to what extent these populations overlap. Sources for this analysis include the official websites of the two tests, various university, professional organisations, and government websites which explicitly state the use of results from these two tests[3].

PTE-Academic is an 'English language test for international study and immigration' (see PTE website). On 16 December 2019, Pearson released this statement: 'PTE Academic is one of the fastest growing products in Pearson and is a strategic growth priority for the company, posting 30% growth in test volumes last year. The test is already accepted by the Australian and New Zealand governments for all visa applications. It is also accepted by 100% of Australian, New Zealand and Irish universities, 98% of UK universities, and more than 2,000 academic programs in the USA.' (Source: https://pearsonpte.com/articles/pearson-awarded-government-commercial-agreements-to-provide-test-of-english-for-people-applying-to-work-or-live-in-the-uk/)

'IELTS is the high-stakes English test for study, migration or work'. According to its official websites:

> The International English Language Testing System (IELTS) is designed to test the English language abilities of non-native speakers who plan to study or work where English is the language of communication. IELTS is accepted by over 10,000 organisations around the world, and more than 3 million IELTS tests were taken worldwide in the last year. Organisations that accept IELTS results include:
>
> - all universities and the vast majority of education providers in Australia, New Zealand and the UK (and most in Canada)
> - more than 3,000 institutions in the US (including Ivy League universities)
> - immigration authorities in Australia, Canada, New Zealand and the UK
> - professional registration bodies worldwide, covering areas such as accounting, engineering, law, medicine and nursing
> - a wide range of employers from sectors such as banking and finance, government, construction, energy and natural resources, aviation, health and tourism

- universities in non-English speaking countries where English is the language of instruction.

It is evident that the two tests target similar populations of test-takers and test-score users. Among these, the three main populations of both tests are university and school students planning to study in EMI contexts, migrants to English-speaking countries, and professionals seeking registration with professional bodies (see also Section 5.3: Inferences). IELTS distinguishes its General Training and Academic tests:

- IELTS General Training is for 'People who are going to English-speaking countries for secondary education, work experience or training programs, or migration to Australia, Canada, New Zealand and the UK.' (See IELTS website.)
- IELTS Academic is for 'People who are applying for higher education or professional registration in an English-speaking environment.' (See IELTS website.)

However, the boundaries of the two specific populations may not be so clear-cut as described by IELTS. Many people 'who are going to English-speaking countries for secondary education, work experience or training programs, or migration to Australia, Canada, New Zealand the UK' are taking IELTS Academic rather than IELTS General Training.

The overlap in the targeted test-taker populations between PTE and IELTS has led to an increasing number of people taking both PTE and IELTS. The three interviewees have all taken IELTS and then PTE several times. We present their test-taking experience and scores achieved below.

Interviewee 1 took IELTS three times at the end of 2015 and January 2016, achieving grades ranging from 6.0 to 7.0. In order to obtain evidence/certificate of her 'Superior English' (at least PTE=79, or IELTS 8.0 in all four skills) as required by the Australian Department of Home Affairs, she then took PTE 10 times within half a year by the end of 2018. She said it was because there was no way that she would be able to achieve IELTS 8.0 that made her decide to take PTE. The first time she took PTE without much preparation, and her PTE overall score was similar to her IELTS score of 6.0; but her PTE scores in the final few tests reached 79 or above, except for the Listening which was just 1 point short of 79. It then reached a kind of plateau as she said. This kind of significant improvement in test scores is also observed in Barkaoui's (2019) study on PTE repeaters' changes in their Writing scores after repeatedly taking the test within five months, as he wrote: 'The findings of this study suggest that test-takers who repeat the test within five months tend to exhibit some significant gains in their writing scores' (p.22).

Interviewee 2 took IELTS several times. The first time, she achieved IELTS 6.5 (Listening=7.0, Reading=7.0, Writing=6.0, Speaking=5.5). She took her final IELTS test just before her A-Level results were announced in summer 2019. As she was a bit concerned that her IELTS score report might not be available by the deadlines set by the universities she chose as Firm and Insurance[4], she decided to take PTE because test results would be available within five business days. She took PTE twice. The first time, her overall score (62) was similar to the IELTS scores she previously achieved; however, she was awarded the full mark (90) for Speaking. She took PTE again, at the shortest interval possible between two tests (five days). Her overall score this time increased drastically (Overall=82; Listening=77, Reading=77, Writing=81), with Speaking remaining the same at full mark (90). At around the same time, she received the score report of her final IELTS test, which was 7.0. The difference between her IELTS and PTE scores was quite big if we compared them using the Pearson's equivalence table (see Table 12). Her PTE overall score of 82 would mean that she had an IELTS score very close to 8.5, or at C2 level of CEFR. However, her official score of IELTS taken around the same time as she took PTE was 7.0.

4. Applicants to undergraduate programs in the UK can make their first choice of offers as Firm and their second choice as Insurance if they have two offers.

Interviewee 3 took IELTS several times, aiming to reach 7.0 for all sections to meet the entry requirements of her Firm choice for her undergraduate study. In July 2019, she achieved IELTS Overall=6.5 (Listening=7.0, Reading=6.5, Speaking=6.5, Writing=6.0). She prepared for PTE for about two weeks and took the test on a Sunday at the end of June 2020. She received her test results on the same night, with the overall score of 71 (Listening=68, Reading=68, Speaking=69, Writing=73) and a wide range of scores of enabling skills (Grammar=76, Oral Fluency=79, Pronunciation=73, Spelling=75, Vocabulary=49, Written Discourse=90). She thinks that the IELTS Writing score is the most challenging for her to improve, as shown in her lower score in Writing than other sections; so she was very pleased that she achieved the full mark in PTE Written Discourse (90) and her highest score in PTE was also in Writing (73), in contrast to her lowest score in IELTS Writing (6.0).

5.2 Constructs and measurement characteristics/conditions

Test content includes all test materials or elements of the tasks presented to candidates to elicit a response. Categories for this analysis can be found in the context validity ('theory-based validity') sections of Weir (2005) and various later publications utilising the socio-cognitive model (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011; Taylor & Weir, 2012; Weir et al, 2013); under 'task demands' and some categories of 'task setting' (which clearly overlap with 'measurement characteristics/conditions' in Kolen and Brennan's terms). Two key questions in relation to test content that we should consider: (a) What are the features that characterise the content of each test? (b) To what extent do these two lists of features overlap?

An understanding of test constructs is informed most by test content analysis. However, we should also consider definitions or statements made by the test providers about the constructs of their tests, and identify any discrepancy between definitions or statements and extrapolations from the content analysis. For this type of analysis, we aimed to answer two questions: (a) How is the construct defined, explicitly and implicitly, for each test? (b) To what extent do the constructs of the two tests overlap?

'Measurement characteristics' is also called 'conditions of measurement' (Kolen, 2007, p.33). For this aspect of analysis, we aimed to answer two questions: (a) What are the measurement characteristics of each test? (b) To what extent are the measurement characteristics of each test similar? These include elements under the control of the test provider (e.g. instructions, timing, layout of venue) and elements beyond the control of the test provider (e.g. the reasons candidates take the test, and their intensive test preparation activities[5]). Various elements of the 'context validity' and 'scoring validity' sections of Weir (2005) and the subsequent publications are relevant and useful guidance on what is to be analysed. As such, there is a significant overlap between the three areas: contents, constructs and measurement characteristics/conditions, as constructs are manifested in test contents which include measurement characteristics/conditions under the control of test providers. We therefore decided to combine the three areas into one, to make the report less repetitive on one hand, and to ensure that readers can get a more comprehensive overview of test contents, constructs and measurement characteristics/conditions in one place, on the other hand.

Following, we report the degrees of similarity between IELTS and PTE in their assessment of the four skills, in the order of Speaking, Writing, Listening, and Reading.

5. It is also important to consider test preparation as an integral part of content and construct analyses, because test preparation may well alter/influence the extent to which test content and construct are actually operationalised as intended by test providers. Coachability of the tests is highly relevant to impacts of test preparation on test construct (Yu, He, Rea-Dickins, et al 2017). However, it is beyond the scope of this report to compare the different test preparation strategies and approaches taken by the candidates for IELTS and PTE, although data from the three interviewees who prepared for the two tests in different ways do provide some glimpses into the complex nature of the relationships between test construct and test preparation.

5.2.1 Speaking

This section reports the degrees of similarity in assessment of Speaking ability between the two tests. Table 1 and Table 2 present an overview of the two Speaking tests.

The IELTS Speaking test takes about 11–14 minutes. It has three parts: initial exchange, long-turn monologue on a given familiar topic, and follow-up discussion on the same topic. The IELTS Speaking test is conducted in face-to-face mode (whether physically in the same room, or remotely via a video call), either before or after the written test. The candidate's spoken monologue and subsequent discussion are recorded. Candidates are given an overall score of 1–9 by a certified examiner according to the band score descriptors which are publicly available to test-takers (see https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en).

Table 1: *Overview of IELTS Speaking tasks*

Item type	Item features	Skill focus	Enabling language skills or sub-scores (as defined in rating scales)
Initial exchange	4–5 minutes The examiner asks general questions about the candidate on a range of familiar and common topics related to personal experience and daily life, such as hometown, family, work, free-time, studies, and interests, designed to encourage spoken interactive communication.	SPEAKING	Fluency and Coherence Lexical Resource Grammatical Range and Accuracy Pronunciation
Long-turn monologue on a given topic	4 minutes including preparation The candidate is given a card which asks them to talk about a particular topic, then has 1 minute to prepare before speaking for up to 2 minutes. The examiner will then ask one or two questions on the same topic.		
Follow-up discussions	4–5 minutes, follow-up discussion The candidate is asked further questions about the topic in Part 2. These involve discussion of more abstract ideas and issues.		

The PTE Speaking section is assessed via five different question types, and each question is individually timed by the computer system (see Table 2). The Speaking test is a component in Part 1 (Speaking and Writing). Although we know the total time allocated for Speaking and Writing is 77–93 minutes, it is not clear how many minutes are actually allocated for Speaking. As some test-takers may be given two essay-writing tasks (each 20 minutes) and one 'summarise written text' task (10 minutes), while others are given one essay-writing task and two 'summarise written text' tasks, it is difficult to know how many minutes exactly are allocated to the Speaking test because the number of items allocated to candidates for each question type is not fixed; however, our best guess would be 27–53 minutes, with an average of 30–35 minutes (see also Pearson, 2019b).

Table 2: *Overview of PTE Speaking tasks*

Item type	Item features	Skill focus	Enabling language skills or sub-scores (as defined in rating scales)	No. of items in one test
Part 1.1 Personal introduction	25 seconds to read the prompt and prepare for the answer 30 seconds to record response once	Not scored, but sent to institutions selected by candidates	N/A	N/A
Part 1.2 Read aloud	Read aloud a text of up to 60 words presented on screen 30–40 seconds to prepare, depending on the length of text 30–40 seconds to read aloud Response recorded once only 3 seconds of silence triggers the recording being stopped	Reading + Speaking	(Content) Oral fluency Pronunciation	6–7
Part 1.3 Repeat sentence	After listening to a recording of a sentence once, repeat the sentence (3–9 seconds)	Listening + Speaking	(Content) Oral fluency Pronunciation	10–12
Part 1.4 Describe image	Describe an image presented on screen in detail 25 seconds to study the image and prepare for response Candidates can take notes on an erasable note-board booklet on screen to write down key ideas, phrases and explanatory details. Time to answer: 40 seconds	Speaking	(Content) Oral fluency Pronunciation	6–7
Part 1.5 Retell lecture	After listening to or watching a lecture on an academic subject, retell the lecture in candidate's own words 3 seconds before the audio starts, an image (as a graph, with or without brief written descriptions) appears on screen Lecture length: up to 90 seconds Time to answer: 40 seconds (+10 seconds to prepare)	Listening + Speaking [Reading skills are only used to read the task instructions and the brief written descriptions of the image; Reading is not assessed]	(Content) Oral fluency Pronunciation	3–4
Part 1.6 Answer short question	Listen to a question, answer with a single word, a few words Prompt length: 3–9 seconds Time to answer: 10 seconds	Listening + Speaking	Vocabulary	10–12

All the PTE Speaking tasks use texts or images of academic or academic-like nature, though of general rather than complex academic content knowledge. In this sense, PTE Speaking has a stronger academic orientation than IELTS in all these tasks (see also similar observations made by Nakatsuhara et al, 2018). However, the academic nature of the three shorter tasks in PTE Speaking (Read aloud, Repeat sentence, Answer short question) is less prominent than the other two Speaking tasks (Describe image, Retell lecture) which provide more substantial visual and/or audio input and require more extended responses from candidates. Take 'repeat sentence' as an example, as van Moere (2012) explained, repeating meaningful sentences, a kind of elicited imitation task in Versant English Test (also owned by Pearson and uses similar automated scoring system), can measure consistently the psycholinguistic processes (e.g. automaticity)

though not the communicative or interactional process of spoken communication. To some extent, as he argued, sentence repetition task can measure fluency and accuracy but not complexity (see also Yan et al, 2016, for a review of studies on elicited imitation tasks).

The IELTS Speaking test is shorter in time (11–14 minutes) and involves interactions between candidate and examiner, and the candidate's performance is more likely to be co-constructed between candidate and examiner. PTE is longer in time, in the range of 27–53 minutes with an average of 30–35 minutes, and is almost twice as long as the IELTS Speaking test time, but does not involve any interaction with a real person.

Both IELTS and PTE Speaking tasks involve candidates' listening, reading and speaking skills. In IELTS, candidates need to listen to instructions from, and interact with, the examiner in all three tasks, and they also need to read the instructions for the long-turn monologue task. Candidates' listening and reading skills would potentially influence their Speaking scores, however, no Listening or Reading score is derived from the Speaking test. In other words, the IELTS Speaking test contributes to only the Speaking score of the test.

The PTE Speaking test is more integrated than IELTS in terms of language skills being assessed and reported. It uses more types of assessment tasks, and is more tightly controlled in terms of time allocation for each question type. PTE Speaking also contributes to the measurement of Listening through its three tasks: 'repeat sentence', 'retell lecture' and 'answer short question'; as well as the measurement of Reading through the 'read aloud' tasks (reading silently texts up to 60 words for 30–40 seconds as a preparation for 'read aloud'). The more substantial Listening and Speaking integrated task is 'retell lecture', which involves listening to a lecture of an academic nature, often naturally occurring and unscripted (or recorded in such a way that it looks like unscripted), sometimes with background noise.

The questions that IELTS examiners ask at the 'initial exchange' stages are probably the closest to the 'answer short question' tasks in PTE. However, answer short question tasks in PTE are less predictable, because some of them do require academic knowledge, while IELTS 'initial exchange' questions are personal and common. However, there is no publicly available information on how much 'repeat sentence', 'retell lecture' and 'answer short question' each contributes to the measurement of Speaking and Listening scores, nor is it clear to what extent 'read aloud' contributes to the measurement of Speaking and Reading.

'Describe image' is the only PTE Speaking task that measures Speaking only. This PTE Speaking task is similar to the IELTS Academic Writing Task 1 (see Yu, He & Isaacs, 2017; Yu, Rea-Dickins & Kiely, 2011) in terms of the use of images as task prompts. However, given more limited time allocated for preparation (25 seconds) and response (40 seconds), what can be orally presented by PTE candidates is more limited than what IELTS candidates can write within the recommended 20 minutes in IELTS Academic Writing Task 1.

IELTS Speaking performance is assessed by human examiners holistically on four dimensions: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation, on a scale of 0–9. For all the PTE Speaking tasks other than the last one ('answer short question'), content, oral fluency and pronunciation are the main focus of assessment. An automated scoring system using Pearson Ordinate technology is used to measure the three areas: content, oral fluency, and pronunciation independently, on the basis that 'automated scoring allows individual features of a language sample (spoken or written) to be analysed independently, so that weakness in one area of language does not affect the scoring of other areas.' (Pearson, 2019a, p.3).

Oral fluency and pronunciation are considered 'enabling skills' and reported with sub-scores which contribute to the Speaking score. Oral fluency and pronunciation are assessed on a scale of 0–5 (Pronunciation: 5=native-like, 4=advanced, 3=good, 2=intermediate, 1=intrusive, 0=non-English; Fluency: 5=native-like, 4=advanced, 3=good, 2=intermediate, 1=limited, 0=disfluent). However, except for the 'answer short question' tasks where content is clearly measured as vocabulary, content in other tasks with different kinds of input (lecture, image, sentence to repeat or to read aloud) is measured on a scale of 0–3 or 0–5, depending on the tasks (see *PTE Academic Score Guide for Test Takers*, v.12, 2019).

The assessment criteria for 'content' vary, dependent on the nature of the PTE tasks. For 'read aloud' and 'repeat sentence', the focus is on counting the number of correct words in response, and for 'answer short question' it is the appropriacy of one or a few words in response. However, for 'retell lecture' and 'describe image', 'ideas' is the main focus of assessment, for example, whether all the main points/ideas of the lecture or image are included in a response, whether ideas (main idea and details) are presented logically, and whether possible conclusions or implications of the lecture are mentioned. In all the PTE Speaking tasks, content of candidates' responses is largely predetermined by the task prompts. In other words, there is less variation in what candidates can say in their responses to the tasks, compared to the three IELTS Speaking tasks.

It is evident that IELTS and PTE Speaking tasks vary to a large extent, in terms of features of task input and conditions, expected responses, and assessment criteria and methods. In PTE, the number of tasks for each question type is not fixed for every candidate, and we do not know the weighting of each task, nor the weighting of each skill assessed in the integrated Speaking tasks. In IELTS, there is no explicitly allotted percentage weighting of 'initial exchange', 'long-turn monologue' and 'follow-up discussion'; IELTS candidates' performance is assessed holistically across the three parts.

As there is little overlap in task features and assessment methods between PTE and IELTS, it is difficult to compare task by task in detail. Therefore, the overall evaluation of the linguistic and cognitive demands of the two Speaking tests had to be made holistically, rather than a precise multiplication of task features by the precise number of such tasks and weightings of the tasks. We used a scale of 1–5 (0.5 point was also used) to identify the degrees of similarity between the two tests. The scale of 1–5 represents only a broad-brush comparison. A task feature given a 5 point does not necessarily mean it has the maximum level of challenge or demand, it simply means it is most demanding or challenging, holistically and comparatively judged against other tasks in the two tests. It should also be noted the values of these numbers are local and specific to the comparison of assessment of a particular skill. In other words, a 2 in Speaking may not be exactly equal to 2 in Writing, in their values.

Table 3 summarises the findings of the analysis, using Weir's (2005) socio-cognitive framework, on the linguistic and cognitive processing demands of each question type, taking into consideration features of task prompts and expected responses such as skill focus, domain coverage (academic vs general content), number of tasks, task length, test length, discourse mode, lexis and grammar, content knowledge, topic familiarity, cultural specificity, nature of information, presentation (verbal, visual, textual input), accent and background noise in input[6], intended speaker/listener relationship, preparation time, and response time.

6. PTE recordings in the 'retell lecture' task include non-native English speaker accents and occasionally background noise. It can increase authenticity of tasks as well as the cognitive processing load. As candidates may be starting at a slightly different time, presented with a different number of Speaking tasks, and candidates do not sit far away from one another, it is possible further real-life and real-time 'background noise' from the PTE test centre is introduced.

Table 3: *Summary of linguistic and cognitive processing demands of the Speaking tasks*

	PTE Speaking					IELTS Speaking		
	Read aloud (6–7 items)	Repeat sentence (10–12 items)	Describe image (6–7 items)	Retell lecture (3–4 items)	Answer short question (10–12 items)	Initial exchange	Long-turn monologue	Follow-up discussion
Overall linguistics	3.0	2.0	4.0	5.0	2.5	2.5	4.0	4.0
Overall cognitive	2.5	2.5	4.0	5.0	2.5	2.0	4.0	4.0

It was found that, on average, PTE Speaking is slightly less linguistically demanding than IELTS Speaking in all tasks except 'retell lecture'. Similar to the views of the three interviewees who have taken both IELTS and PTE multiple times, our analysis found that 'retell lecture' is the most linguistically and cognitively demanding among all the Speaking tasks, both in terms of task input and expected response.

In terms of time pressure and consequently cognitive processing load in a short space of time, PTE Speaking is slightly more cognitively demanding than IELTS Speaking. However, as Interviewee 3 said, the linguistic and cognitive demands of the two Speaking tests are rather different and fluid due to not only candidates' Speaking abilities but also the approaches or strategies they would take to deal with the tasks. For PTE Speaking, as she said, her strategies were to keep talking to the computer as 'fluently' or quickly as possible to bump up her scores in the enabling skills of 'oral fluency' and 'pronunciation'. In the IELTS Speaking test, she had to consider the reaction in real-time from the examiner, which can be intimidating to candidates. Similar test-taking and preparation strategies were also reported by other PTE candidates in Knoch et al's (2020) qualitative study using semi-structured interviews with around 60 participants.

5.2.2 Writing

This section reports the comparisons between PTE and IELTS Writing tasks. Table 4 and Table 5 present an overview of IELTS and PTE Writing tasks, respectively.

Table 4: *Overview of IELTS Writing tasks*

Item type	Item features	Skill focus	Enabling language skills or sub-scores (as defined in rating scales)
Task 1	Write a descriptive essay, in a formal style, based on information visually given in a graph/table/diagram, etc. Recommended time: 20 minutes Response: 150 words Weighting: half that of Writing Task 2, contributing to the Writing band	Writing	Task achievement (Content) Coherence and cohesion Lexical resource Grammatical range and accuracy
Task 2	Write an argumentative essay, in an academic or formal (or semi-formal/neutral) style, on a given point of view, argument or problem Recommended time: 40 minutes Response: 250 words Weighting: twice that of Writing Task 1, contributing to the Writing band	Writing	Task response (Content) Coherence and cohesion Lexical resource Grammatical range and accuracy

Table 5: *Overview of PTE Writing tasks*

Item type	Item features	Skill focus	Enabling language skills or sub-scores (as defined in rating scales)	No. of items in one test
Part 1.7 Summarise written text	Prompt: a text of up to 300 words Response time: 10 minutes Write a summary of the text, in a full, single sentence of up to 75 words (word count is shown as candidates type their response)	Reading + Writing	(Content) (Form, i.e. word count) Grammar Vocabulary	2–3
Part 1.8 Write essay	Prompt: 2–3 sentences Response time: 20 minutes Write a 200–300-word argumentative essay on a given topic (word count is shown as candidates type their response)	Writing [note: reading skills are only used to read the instructions and the prompt]	Content) (Development, structure, and coherence) (Form, i.e. word count) (General linguistic range: ideas) Grammar usage and mechanics Spelling Vocabulary range Written discourse	1–2
Part 2.1 Reading & Writing: Fill in the blanks	Prompt: text up to 300 words appears on screen with several gaps Choose a word from a drop-down list of four options (of similar spelling, but completely different meaning) for each blank	Reading + Writing	(ability to use contextual and grammatical cues to identify words that complete a reading text)	5–6
Part 3.1 Summarise spoken text	After listening to a recording, write a 50–70-word summary Prompt: a short lecture, 60–90 seconds Response time: 10 minutes Note-taking allowed while listening to the recording	Listening + Writing	(Content) Form Grammar Spelling Vocabulary	2–3
Part 3.3 Listening & Writing: fill in the blanks	A transcript of a recording appears on screen with several gaps; after listening to the recording, type the missing word in each gap Prompt length: 30–60 seconds Candidates have 7 seconds to skim the transcripts before the audio starts Note-taking allowed while listening to the recording	Listening + Writing	(Content: correct word)	2–3
Part 3.8 Write from dictation	After listening to a recording of a sentence, type the sentence Prompt length: 3–5 seconds	Listening + Writing	(Content: correct words)	3–4

Note: The brackets in the Enabling skills/score column mean that they are the focus of assessment, but not reported as an enabling skills score.

Both IELTS and PTE Writing tasks target a range of proficiency levels, using texts (verbal or textual/visual) of varying degrees of difficulty. They are designed to elicit written responses of different genres and lengths. In the case of IELTS, both Writing tasks require an extended discourse (150–250 words). In PTE, candidates produce a mixture of written responses of different lengths, from selecting or typing a single word (in the two 'filling in blanks' tasks with Reading and Listening inputs respectively), repeating a short sentence (in 'write from dictation'), producing a more extended sentence (up to 75 words in the 'summarize written text'), a few sentences (50–70 words for 'summarize spoken text'), to a more extended discourse (300 words for 'write essay').

IELTS candidates are given 60 minutes to complete the two writing tasks, and they are recommended to write for 20 minutes for Task 1 and 40 minutes for Task 2, although how exactly candidates allocate their time is entirely up to themselves. The total time allocated for the PTE Writing test does not seem to be fixed, although each task is individually timed. One interviewee said she was allocated two 'summarize written text' tasks (10 minutes each) and one 'write essay' task (20 minutes) in one sitting of the test, and one 'summarize written text' task and two 'write essay' tasks in another sitting.

As seen in Table 4 and Table 5, the only two comparable tasks between the two tests in terms of features of task prompt and expected response are the PTE 'write essay' task and IELTS Task 2; both tasks give candidates a topic/argument/problem for them to write a persuasive/argumentative essay (though with a different requirement on length).

PTE uses more task types than IELTS (six compared to two) to assess writing, and PTE is also more 'integrated' in its assessment methods, i.e. integrating assessment of writing with other language skills (Yu, 2013a). The six PTE tasks that contribute to the assessment of writing are spread across the three parts/papers of the test: Part 1 (Speaking & Writing), Part 2 (Reading) and Part 3 (Listening).

A significant difference between PTE and IELTS is the use of two types of summarisation tasks in PTE (Yu, 2013b) – 'summarize written text' of up to 300 words into a single, full sentence of up to 75 words, and 'summarize spoken text' (i.e. a short academic lecture of 60–90 seconds) into 50–70 words in writing. The 'summarize spoken text' in writing (from the Listening paper) shares the same features of task prompt with the 'retell lecture' task in the Speaking paper. The tight time control makes the summarisation tasks more cognitively challenging. The two 'fill in blanks' tasks (read to fill in blanks, listen to fill in blanks), however, are indirect assessment of writing, as they mainly assess candidates' knowledge of single words (i.e. 'vocabulary') with contextual and grammatical cues. The 'listen to fill in blanks' task is like a partial-dictation test. Candidates are required to fill in blanks with single words while listening to a recording. In the 'read to fill in blanks' task, candidates are given a passage with blanks to fill in by choosing one word from a drop-down list of four words for each blank. This task mainly assesses candidates' lexical knowledge, rather than their real writing ability as such. The two 'fill in blanks' tasks are designed to assess Reading and Listening respectively (see Table 5). The 'write from dictation' task (from the Listening paper), which requires candidates to type a sentence they have just heard, is basically the written version of the 'repeat sentence' task (from the Speaking paper) as they share the same kind of task input. There is no publicly available information on the weightings of each skill in the tasks that contribute to the assessment of more than one communicative or enabling skill (see Section 6: Discussions and Conclusion).

Weir's (2005) socio-cognitive framework was used to guide the analysis of the linguistic and cognitive processing demands of each Writing task, from the perspectives of features of task prompts and expected responses, taking into consideration features such as: skill focus, domain coverage (academic vs general content), number of tasks, task length, test length, discourse mode, content knowledge, topic familiarity, cultural specificity, nature of information, presentation (verbal, visual, textual input), intended writer/reader relationship, and preparation and response time.

Overall, the PTE Writing test was considered moderately more demanding, both linguistically and cognitively, as it requires candidates to complete a range of independent and integrated tasks of different kinds of inputs and response formats under tight time control for each question. Table 6 presents a summary of the linguistic and cognitive processing demands of each Writing task in the tests. However, it should be noted, as in our analysis of PTE Speaking tasks, that there is also a lack of information on how much each task contributes to the assessment of writing. There is no information on the weighting of each skill in the writing tasks that are supposed to measure both writing and another skill (see Pearson 2019b, Academic Score Guide for Test Takers, v.12).

Table 6: *Summary of linguistic and cognitive processing demands of the Writing tasks*

	IELTS		PTE					
	Task 1 (graph-based) 1 item	Task 2 (topic-based argu-mentative essay) 1 item	Summarize written text (2–3 items)	Write essay (1–2 items)	Reading & Writing: fill in the blanks (5–6 items)	Summarize spoken text (2–3 items)	Listening & Writing: fill in blanks (2–3 items)	Write from dictation (3–4 items)
Linguistic demand	4.0	4.0	5.0	4.0	4.0	5.0	2.5	2.5
Cognitive demand	4.0	4.0	5.0	4.0	3.5	4.5	2.5	2.5

The interview data, as well as the high PTE scores that Interviewees 2 and 3 achieved, seemed to contradict the finding from the textual analysis of the writing tasks that PTE Writing is moderately more challenging than IELTS. Interviewees 2 and 3 both had their lowest IELTS scores in Writing; but the highest scores in PTE Writing (or the second highest in the case of Interviewee 2 who had a full mark in PTE Speaking). This sharp discrepancy suggests that PTE and IELTS may be assessing different kinds of writing skills. In the case of Interviewee 3, she achieved a full mark in 'written discourse'; but she said she was struggling to achieve a high score in IELTS Writing.

5.2.3 Listening

The IELTS Listening test has four sections, with 10 questions of different types in each section. The recordings could include a range of accents, including British, Australian, New Zealand, American and Canadian. For the first two sections, they are of everyday social contexts: a conversation between two speakers (for Part 1), and a monologue, e.g. a speech about local facilities (for Part 2). Recordings for Parts 3 and 4 are from situations set in educational and training contexts. For Part 3, the recording is a conversation between two (main) speakers, e.g. a conversation between two university students, guided by a tutor. Part 4 is a monologue on an academic subject. As such, Parts 1 and 2 are relatively easier than Parts 3 and 4.

The questions are designed in such a way that they are in the same order as they appear in the recording, which to some extent provides scaffolding to candidates and makes the tasks less dependent on short-term memory and therefore less cognitively demanding.

Candidates are given some time to read/skim the questions before the audio starts, which gives listeners some time to get a sense or predict what the recording might be about before listening to the recordings. This arrangement helps to further reduce the cognitive demands of the tasks. The recordings are heard only once. At the end of each part within a section, test-takers have some time (about 30 seconds) to check their answers before moving onto the next part of the section. At the end of the Listening test, they have 2 minutes to check their answers. The time for the Listening test is between 30–34 minutes. Although they do not contribute to the Reading score, reading skills are essential for successful performance in the IELTS Listening test because candidates do need to read the task prompts quickly (sometimes quite substantial chunks of text in Part 4) before the recording starts to play; and they need to read much more closely while the recording is being played and they are filling in the blanks or choosing an answer on the go.

A variety of question types are used, including multiple choice (single answer with three options; multiple answers with several options to choose from), matching, labelling (e.g. a plan, map, or diagram) from a list of options, and completion of a form, note, table, flow-chart, a summary, or a sentence. Some questions require candidates to write their answers (fill in blanks), mostly in single words, rather than choosing from a list of words. Each answer is worth 1 mark, regardless of the difficulty level of the recordings or question type. The raw score out of 40 is converted to a band score: 16/40=Band 5; 23/40=Band 6; 30/40=Band 7; 35/40=Band 8.

Table 7 presents an overview of the PTE Listening tasks. PTE Listening is assessed in two papers (Part 1: Speaking & Writing; Part 3: Listening). In Part 1 (see Table 2, and also repeated in Table 7), there are three integrated Listening/Speaking tasks ('repeat sentence', 'retell lecture' and 'answer short question', see Section: 5.2.1: Speaking). In the three integrated Listening/Speaking tasks in Part 1, 'Listening' is the input for the Speaking tasks and contributes to the assessment of Listening as well. Part 3, the separate Listening paper, has eight question types (see Table 7). Of the eight question types, three also assess Writing (see Section 5.2.2: Writing) and two also assess Reading (see Section 5.2.4: Reading). The weightings on Listening assessment in these integrated tasks, however, are unknown.

Three further question types (multiple-choice question: multiple answers; multiple-choice question: single answer; and selecting missing words) assess Listening only. PTE allocates 45–57 minutes to the Listening paper, in addition to the three Listening/Speaking tasks in the first part of the test. Candidates listen to audio or video input and questions once only. They can adjust the volume of the recordings and take notes on the erasable note-board area on screen.

Table 7: *Overview of PTE Listening tasks*

Item type	Item features	Skill focus	Enabling skills/scores	No. of items in one test
Part 1.3 Repeat sentence	After listening to a recording of a sentence once, repeat the sentence (3–9 seconds)	Listening + Speaking	(Content) Oral fluency Pronunciation	10–12
Part 1.5 Retell lecture	After listening to or watching a lecture on an academic subject, retell the lecture in candidate's own words 3 seconds before the audio starts, an image (as a graph, with or without brief written descriptions) appears on screen Lecture length: up to 90 seconds Time to answer: 40 seconds (+10 seconds to prepare)	Listening + Speaking [Reading skills are only used to read the task instructions and the brief written descriptions of the image, but Reading is not assessed]	(Content) Oral fluency Pronunciation	3–4
Part 1.6 Answer short question	Listen to a question; answer with a single word or a few words Prompt length: 3–9 seconds Time to answer: 10 seconds	Listening + Speaking	Vocabulary	10–12
Part 3.1 Summarize spoken text	After listening to a recording, write a 50–70-word summary Prompt: a short lecture, 60–90 seconds Response time: 10 minutes	Listening + Writing	(Content) Form Grammar Spelling Vocabulary	2–3
Part 3.2 Multiple choice, multiple answer	After listening to a recording, answer a multiple-choice question on the content or tone of the recording by selecting more than one response Tests: both main idea and details Candidates have 7 seconds before the recording begins to preview the question and options Prompt: 40–90 seconds	Listening		2–3
Part 3.3 Listening & Writing: fill in the blanks	A transcript of a recording appears on screen with several gaps; after listening to the recording, type the missing word in each gap Prompt length: 30–60 seconds Candidates have 7 seconds to skim the transcripts before the audio starts	Listening + Writing	(Content: correct word)	2–3
Part 3.4 Highlight correct summary	After listening to a recording, select the paragraph that best summarises the recording Prompt length: 30–90 seconds	Listening + Reading	(ability to comprehend, analyse and combine information from a recording and identify the most accurate summary of the recording)	2–3

Part 3.5 Multiple choice, single answer	After listening to a recording, answer a multiple-choice question on the content or tone of the recording by selecting one response; questions assess 'main idea', 'details', 'inference', 'author's purposes', etc. Prompt length: 30–60 seconds Candidates have 5 seconds to skim the question and answer options before the recording begins to play	Listening		2–3
Part 3.6 Select missing word	After listening to a recording, select the missing word (or a group of words) that completes the recording from a list of options – to predict what the speaker will say based on contextual cues Prompt length: 20–70 seconds	Listening		2–3
Part 3.7 Highlight incorrect words	The transcript of a recording appears on screen Candidates have 10 seconds to skim the transcript (but it's not possible to read word-by-word) before the recording begins While listening to the recording, identify the words in the transcript that differ from what is said Candidates select the wrong words as the text is read Prompt length: 15–50 seconds	Listening + Reading		2–3
Part 3.8 Write from dictation	After listening to a recording of a sentence, type the sentence Prompt length: 3–5 seconds	Listening + Writing	(Content: correct words)	3–4

Using Weir's (2005) socio-cognitive framework, we compared the linguistic and cognitive demands of the Listening tasks, taking into consideration features of recordings, questions and options, and expected responses, such as: speech rate, accent and background noise in the recordings, length of recording, number of recordings, domain coverage (academic vs general content), content and topic familiarity, cultural specificity, discourse mode, lexis, grammatical and syntactic complexity, nature of information, presentation (verbal, visual, textual), intended listener/speaker and listener/writer relationships (for PTE integrated listening/speaking, and listening/writing tasks only), number of tasks, task length, preview and preparation time, response time, response mode and method, and test length.

Overall, both the IELTS and PTE Listening tests are tightly controlled in time for each question. However, the number of questions that PTE candidates complete is not fixed. In IELTS, each candidate completes the same number of questions and listens to the same recordings.

Both IELTS and PTE include speeches of an academic nature. However, IELTS includes half of its questions (i.e. 20 in the first two sections) based on recordings of a more general nature. IELTS recordings are longer than PTE recordings if we compare each single recording, but PTE uses more task types and more recordings. The PTE Listening test is longer than the IELTS test. IELTS has a relatively slower than normal speech rate, and a slower speech rate than PTE recordings. PTE recordings sound more natural. PTE recordings sometimes include background noise or/and some irrelevant speech. PTE speech rates also vary within a recording.

IELTS provides more textual support, in the form of texts and sometimes visual aids (e.g. diagrams, tables and maps) to test-takers, than PTE does. The textual support, coupled with the questions being presented in the order of content of the recordings, helps reduce the cognitive load of the IELTS Listening tasks, especially those tasks that require candidates to fill in blanks with specific details. However, PTE uses this kind of input for assessment of Reading at the same time.

The 'listen to fill in the blanks' task in PTE (Part 3.3, Listening and writing: Fill in the blanks) is easier than IELTS's 'fill in blanks' task because the former task presents the whole scripts to candidates. In IELTS, candidates are presented with only partial scripts in an outline form, which makes it more challenging to complete than PTE's 'listen to fill in the blanks' task.

Multiple choice (single answer) questions are the question type which is most comparable between IELTS and PTE. However, differences even between the two are also quite apparent as IELTS uses three options while PTE uses more options. Various research studies have confirmed the efficiency of using three options for multiple-choice questions, but it is a quite debatable area (see, for example, Shizuka et al, 2006). The PTE multiple-choice questions with multiple answers are more challenging than multiple-choice questions with a single answer.

PTE uses more integrated Listening tasks than IELTS. Listening is integrated with Speaking and Writing tasks, to a significant extent. Listening is also integrated, to a smaller extent, with assessment of Reading in two task types (Part 3.7, 'highlight incorrect words'. Test-takers listen to identify incorrect words of the whole scripts presented on screen; and Part 3.4, 'highlight correct summary'. Test-takers listen to identify a correct summary out of four choices). The integration of the Listening/Speaking and Listening/Writing tasks is much higher than that of the two Listening/Reading tasks. Two of the three summarisation tasks in PTE: 'retell lecture' (listen to orally summarise an academic lecture of up to 90 seconds, with 10 seconds' preparation time and 40 seconds' response time) and 'summarize spoken text' (listen to write a summary of an academic lecture of up to 90 seconds, with 10 minutes' response time) involve substantial language production and skills to summarise the main ideas of academic lectures, a feature desirable for tests of English for academic purposes. The third summarisation task in PTE (Part 3.4: 'highlight correct summary'), as a multiple-choice summarisation task (Huhta & Randell, 1996), involves no productive skills, and is therefore an easier task. Wei and Zheng's (2017) analysis of over 5,000 PTE test-takers' performances confirmed that it is indeed the third-easiest task among the 11 Listening tasks (the easiest two being 'repeat sentence' and 'highlight incorrect words'; and the fourth-easiest being 'fill in the blanks'). The two summarisation tasks ('retell lecture', 'summarize spoken text') are more challenging than any of the IELTS Listening tasks, and they are also the most challenging among all PTE Listening tasks (see also Rukthong & Brunfaut, 2020). This was, however, only partially confirmed by Wei and Zheng's (2017) analysis of actual test performance data. They found that 'summarize spoken text' was indeed the most challenging, but 'retell lecture' was the fifth-easiest Listening task – in the middle of the 11 tasks.

Wei and Zheng (2017) identified that 'repeat sentence', 'highlight incorrect words', 'highlight correct summary', and 'fill in the blanks' are the four easiest tasks; and 'write from dictation', 'repeat sentence' and 'select missing word' are the three best predictors of overall listening performance. It was not a surprise that 'repeat sentence', 'highlight incorrect words', 'highlight correct summary' and 'fill in the blanks' were identified as the four easiest tasks because these tasks involve minimal or no language production or transformation from the source materials. However, we were rather surprised to read that the most challenging task, 'summarize spoken text', was not a good predictor of candidates' overall performance in Listening. Instead, they found that it was 'write from dictation', 'repeat sentence' and 'select missing word' that were the best predictors of overall listening performance.

Table 8: *Summary of the linguistic and cognitive demands of the Listening tasks*

	IELTS				PTE										
	1	2	3	4	RS	RL	ASQ	SST	MCQ-M	L&W-F	HCS	MCQ-S	SMW	HIW	WfD
Linguistic demand	2.5	3.0	3.0	4.0	2.0	5.0	2.5	5.0	4.0	2.5	3.0	3.0	2.5	2.0	2.5
Cognitive demand	3.0	3.0	3.5	4.0	2.5	5.0	2.5	4.5	4.0	2.5	3.0	3.0	2.5	2.5	2.5

Notes:

(1) RS=repeat sentence; RL=retell lecture; ASQ=answer short question, SST=summarize spoken text, MCQ-M=multiple choice question: multiple answers, L&W-F=listening and writing: filling blanks; HCS=highlight correct summary; MCQ-S=multiple choice question: single answer; SMW=select missing words; HIW=highlight incorrect words; WfD=write from dictation

(2) As each Section of IELTS Listening paper uses a mixture of different assessment methods, for example multiple choice questions may be used together with fill-in-blanks in a section. Therefore, the holistic evaluation is not based on assessment methods.

We used a scale of 1–5 to make a holistic evaluation of the challenges of each PTE Listening task type (see Table 8 above). Overall, it was found that PTE is slightly more linguistically and cognitively demanding. This finding is different from Taylor and Chan (2015) who considered PTE Listening less demanding than IELTS Listening in most facets of the socio-cognitive framework. As we know, the difficulty level of any assessment task is not simply determined by task features alone; it is ultimately defined by the assessment criteria and the weightings of the tasks in the tests. We do not know how exactly the PTE scoring operates, nor do we know the weightings of each task and each enabling skill in integrated Listening tasks. To a large extent, however, it is probable that these easier tasks in PTE Listening which assess mainly understandings of words at a local level may well cancel out or neutralise the difficulties from the more challenging tasks such as 'summarize spoken text' and 'retell lecture' that require and assess global understanding of the audio inputs. As a result, it is probable that the overall linguistic and cognitive demands of IELTS and PTE are not massively different, at least from the perspectives of scores to be awarded.

5.2.4 Reading

The IELTS Reading section has three reading passages with 40 questions in total, and 60 minutes allocated. Candidates have the freedom to decide how many minutes they spend on each reading passage or question within the 60 minutes allocated. IELTS uses several question types, including multiple choice, True/False/Not Given, answering short question, filling in blanks, and matching, for example:

- identifying specific information and writer's views/claims
- matching specific information, matching headings, matching features, matching sentence endings
- sentence completion, summary completion, note completion, table completion, flow-chart completion, diagram label completion.

Each passage can have a different combination of question types. Each question is worth 1 mark. A raw score out of 40 is converted to the band score, for example, 15/40=Band 5; 23/40=Band 6; 30/40=Band 7; 35/40=Band 8. Reading passages are selected from a variety of sources, including books, journals, magazines and newspapers, suitable for those who are entering for undergraduate, postgraduate programs or seeking professional registration (see for example, Taylor and Chan 2015 on the use of IELTS test results by the General Medical Council in the UK). The passages are in a variety of genres, including narrative, descriptive and argumentative. They may contain non-verbal materials such as graphs or illustrations. The passages are written for a non-specialist audience but overall are largely academic. A simple glossary is provided for any technical term (e.g. behaviourism, humus). Using TextInspector tools, we analysed the texts in Cambridge IELTS Volumes 13 and 14; the reading passages are broadly within B2(+) and C2 levels.

As an example, Table 9 below shows the results of the analysis of all reading passages used in the four tests in Cambridge IELTS Volume 13. However, it should be noted that the difficulty level of the Reading section is not determined entirely by the difficulty level of the reading passages. What questions are asked and how they are asked (in terms of question types) can also influence the difficulty level of the test, to a large extent. For example, the multiple-choice questions (single or multiple answers) in IELTS are presented in the same order as the location the answers appear in the text, which can reduce the cognitive load of the test and make it easier than if the location of the answers were less predictable (see also the similar design in the IELTS Listening test, Section 5.2.3: Listening).

Table 9: *Examples of IELTS Reading passages at CEFR level*

Test	Text	CEFR level	Native speaker academic text %*
1	Tourism New Zealand website	C2	74
	Boredom	B2+	52
	Artificial artists	C1	63
2	Bringing cinnamon to Europe	C2	71
	Oxytocin	C1+	70
	Making the most of trends	D1	85
3	The coconut palm	C2	72
	Baby talk	C1	60
	Harrapan civilisation	C2+	82
4	Cutty Sark: the fastest sailing ship	C1	60
	Saving the soil	C1+	63
	Book review: happiness industry	C2+	77

*Higher % means the reading passage is more academic, written by native speakers.

As shown in Table 10, PTE Reading is mainly assessed through Part 2: Reading paper, which is single timed. As in IELTS, candidates have the freedom to decide how many minutes they would spend on each question within the allocated time for the whole paper (32–40 minutes). Unlike IELTS, however, the number of items that candidates are allocated is not fixed or the same; it generally contains 15–20 items, depending on the combination of items. Five question types are used, including multiple choice (single answer, multiple answers), re-ordering paragraphs, and filling in blanks. The only question types that IELTS and PTE share the largest similarity in assessment of Reading comprehension are the two multiple-choice questions (single answer and multiple answers).

Table 10: *Overview of PTE Reading tasks*

Item type	Item features	Skill focus	Enabling skills/scores	No. of items in one test
Part 1.2 **Read aloud**	Read aloud a text of up to 60 words presented on screen 30–40 seconds to prepare, depending on the length of text 30–40 seconds to read aloud Response recorded once only 3 seconds of silence triggers the recording being stopped	Reading + Speaking	(Content) Oral fluency Pronunciation	6–7
Part 1.7 **Summarise written text**	Prompt: a text of up to 300 words Response time: 10 minutes Write a summary of the text, in a full, single sentence up to 75 words (word count is shown as candidates type their response)	Reading + Writing	(Content) (Form, i.e. word count) Grammar Vocabulary	2–3
Part 2.1 **Reading & Writing: Fill in the blanks**	Prompt: text of up to 300 words appears on screen with several gaps Choose a word from a drop-down list of four options (of similar spelling, but completely different meaning) for each blank	Reading + Writing	(ability to use contextual and grammatical cues to identify words that complete a reading text)	5–6
Part 2.2 **Multiple choice, multiple answer**	After reading a text, answer a multiple-choice question (which statement is true?) on the content or tone of the text by selecting more than one response Prompt: text of up to 300 words	Reading	Penalty for selecting a wrong answer: 1 point off for each wrong answer, until it reaches 0	2–3
Part 2.3 **Re-order paragraphs**	Several text boxes appear on screen in a random order Put the text boxes in the correct order, by selecting text boxes and dragging them across the screen Prompt: text of up to 150 words	Reading	Ability to understand the organisation and cohesion of an academic text If all text boxes are in the correct order, the maximum score points for this question type are given; however, if one or more text boxes are in the wrong order, partial credit scoring applies	2–3

Part 2.4 **Fill in the blanks**	A text appears on screen with several gaps Drag words from the box below the text to fill the gaps; not all words provided are used Prompt: text of up to 80 words	Reading		4–5
Part 2.5 **Multiple choice, single answer**	After reading a text, answer a multiple-choice question on the content or tone of the text by selecting one response Prompt: text of up to 300 words	Reading		2–3
Part 3.4 **Highlight correct summary**	After listening to a recording, select the paragraph that best summarises the recording Prompt length: 30–90 seconds	Listening + Reading	(Ability to comprehend, analyse and combine information from a recording and identify the most accurate summary of the recording)	2–3
Part 3.7 **Highlight incorrect words**	The transcript of a recording appears on screen Candidates have 10 seconds to skim the transcript (but it's not possible to read word-by-word) before the recording begins While listening to the recording, identify the words in the transcript that differ from what is said Candidates select the wrong words as the text is read Prompt length: 15–50 seconds	Listening + Reading	(Correct selection is awarded 1 point; for any wrong selection, 1 point is deducted until it reaches 0)	2–3

One 'fill in the blanks' task assesses Reading only, and the other assesses both Reading and Writing. In the 'Reading & Writing: Fill in the blanks' task, candidates are asked to read a text of up to 300 words with blanks to be filled in, by selecting one word from a drop-down list of four words. Some of these texts are at C2 level, but the task mainly assesses vocabulary, rather than global understanding of the whole text. In the 'Reading: Fill in the blanks' task, candidates are asked to read and fill in blanks in a text of up to 80 words by dragging a word from a box of several words (not all words are used).

Our analysis of the sample texts in PTE websites showed that the texts for the 'Reading: Fill in the blanks' task are relatively easier than the texts for the 'Reading & Writing: Fill in the blanks' task. Although, as found in McCray and Brunfaut (2018), the banked gap-filling task (i.e. 'Reading: Fill in the blanks') was able to differentiate the 28 participants in their study, with evidence of lower-proficiency participants showing more use of lower-level, local processing than higher-proficiency participants according to the participants' eye-movement data, it is not clear to what extent the banked gap-filling task assesses global understanding of the texts. IELTS also uses quite extensively a range of 'fill in the blanks' (or 'completion') tasks in the Reading paper; however, IELTS 'completion' tasks (e.g. sentence completion, summary completion), generally speaking, focus more on assessment of comprehension of content of the source texts (Taylor, 2013) rather than single words as in the 'fill in the blanks' tasks in PTE.

In addition to the five question types in the Reading paper, PTE also uses four integrated skills tasks to assess Reading: two question types in Part 1 paper (Speaking & Writing): 'read aloud' and 'summarize written text'; and two question types in Part 3, the Listening paper: 'highlight correct summary' after listening to a recording, and 'highlight incorrect words' in the transcripts of the recording presented on screen while listening to the recording and reading the transcripts at the same time. However, apart from 'summarize written text' which does involve substantial reading, the other three additional tasks (read aloud a text of up to 60 words, highlight correct summary, highlight incorrect words) can only assess reading minimally. The extent to which 'summarize written text' into one single sentence of up to 75 words can assess candidates' reading or/and writing skills is, nevertheless, dependent on the PTE rating scales, which are not fully transparent. 'Highlight correct summary' after listening to a recording requires only minimal reading of the options (see Section 5.2.3: Listening). 'Read aloud' requires reading a short academic text of up to 60 words (see Section 5.2.1: Speaking). As in the assessment of Speaking, Writing and Listening (see Sections 5.2.1: Speaking, 5.2.2: Writing, and 5.2.3: Listening), a couple of more fundamental questions arise – what is the weighting of Reading in the assessment tasks that assess Reading and another skill, and what is the weighting of each question type in the assessment of Reading? (See further discussion on weighting in Section 6).

Table 11: *Summary of the linguistic and cognitive demands of the Reading tasks*

	IELTS			PTE								
	1	2	3	RA	SWT	R&W-F	MCQ-M	ROP	RD-F	MCQ-S	HCS	HIW
Linguistic demand	4.5	4.5	4.5	3.0	5.0	4.0	4.0	3.5	3.5	4.0	3.0	2.0
Cognitive demand	4.5	4.5	4.5	2.0	5.0	3.5	4.5	3.5	3.5	4.0	3.0	2.5

Notes:

(1) RA=read aloud; SWT=summarize written text; R&W-F=reading and writing: fill in blanks; MCQ-M=multiple-choice questions with multiple answers; ROP=Re-order paragraphs; RD-F=Reading: fill in blanks; MCQ-S=multiple-choice question: single answer; HCS=highlight correct summary; HIW=highlight incorrect words

(2) As each passage of the IELTS Reading paper uses a mixture of different assessment methods, for example, multiple-choice questions may be used together with fill-in-blanks for one reading passage, therefore, the holistic evaluation is not based on one assessment method. Because it is difficult to differentiate the level of challenges of the three passages, I have put the same value for three passages.

The analysis of the linguistic and cognitive demands of the Reading tasks showed that PTE Reading is less demanding than IELTS Academic Reading, although they are very similar in terms of lexical features, readability, topics, rhetoric organisations, presentation, nature of information, discourse mode, content knowledge, and cultural specificity of the source texts, which are broadly between B2+ and C2 levels. IELTS uses a smaller number of much longer texts than PTE (around 1,000 words in PTE vs. close to 3,000 words in IELTS). PTE uses a larger number of short texts. PTE Reading test is shorter than IELTS (some 40 minutes vs 60 minutes).

5.3 Inferences

Inferences (Kolen & Brennan, 2014, p.498) are the conclusions that users of test results are entitled to draw about candidates' ability based on those results and include candidates' own interpretations of their test results and language abilities and any comparisons they would make with reference to other tests they have taken. They can also include statements made by test providers about the link between test results and other external frameworks or tests. However, such 'inferences' should not only be based on the test level claimed by test providers but also on the analyses of content and construct and on any additional information available.

For the purposes of comparing the two tests, we looked at inferences applicable between 4.5 and 8 on IELTS, and 30 and 90 on PTE. There are two key questions to answer: (a) What inferences about candidates, stated or implied, is each test intended to facilitate? (b) To what extent do the two sets of inferences overlap?

As we discussed earlier in Section 5.1: Populations, both tests have largely the same target populations and markets, therefore the inferences about candidates' abilities are intended to facilitate the same kind of decisions: for academic study, immigration, and professional registration. Before we talk about any inferences that users (including test-takers and other score users) could make about test results, it is essential to know how IELTS and PTE report test results.

The IELTS Test Report Form presents the four sub-scores of Listening, Reading, Speaking, and Writing on a scale of 0–9 (including 0.5), based on candidates' performance in the respective papers which have equal weighting in the calculation of the Overall Band Score – an average of the four sub-scores rounded up or down to a half or whole band. If the average of the four components ends in .125, the Overall Band Score is rounded down to the whole band; if the average ends in .25, the Overall Band Score is rounded up to the next half band, and if it ends in .75, the Overall Band Score is rounded up to the next whole band. The Overall Band Score is also presented as a CEFR language proficiency level from A1 to C2.

All the PTE responses, including constructed spoken and written responses, are computer scored, employing 'two automated scoring systems, Intelligent Essay Assessor (IEA) and the Ordinate scoring system (OSS), for its written and spoken sections, respectively' (Wang et al, 2012, p.606). Although PTE uses several integrated assessment tasks (reading and speaking, listening and speaking, reading and writing, listening and writing, or listening and reading), the Score Report presents separately the four communicative/language skills (Listening, Reading, Speaking and Writing) and six enabling skills scores (grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse) on a scale of 10–90. It also reports an overall score, which is not an average of the four communicative/language skills scores, or an average of the six enabling skills scores. There is no publicly available information on how the raw scores are converted to PTE scores, neither is there any information on the weightings of different items and different skills in the integrated assessment tasks (see further discussion on weighting below). The Score Report includes the expiry date of the results (within 2 years of the report date) and shows whether the test results were achieved the first time or not; however, it does not say which time it was if a candidate has taken the test multiple times.

As seen in the score reports, IELTS assumes equal weighting of the four components/ skills to calculate the overall band. In the case of PTE, however, we only know that PTE does not report the average of the four components as the Overall Score, though it is quite close to an average. How exactly an overall score is calculated is unknown to candidates or other score users. A number of questions arise with the relation to weighting, for example:

- What is the weighting of each of the four communicative/language skills scores and each of the six enabling skills scores in the calculation of the overall score?
- What is the weighting of the six enabling skills scores in the calculation of each of the four communicative/language skills scores?
- What is the weighting of each item and each question type in the assessment of a communicative/language skill?
- What is the weighting of each item and each question type in the assessment of an enabling skill?

Another important population is the end users of test scores, e.g. governments and universities. Some users, including for example the Australian Government, use the alignments provided by Pearson in setting IELTS and PTE scores for different levels of English.

Table 12: *PTE and IELTS equivalence as reported by Pearson*

PTE (as now*)	30**	36**	42	50	58	65	73	79	83	86–90
IELTS	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0
PTE in 2009	29**	35**	42	50	58	64	72	79	83	87–90

*PTE Score Guide v12 (October 2019)
** In Score Guide v12, it was 29 for 4.5, and 35 for 5.0; however, on the website as now it is 30 for 4.5, and 36 for 5.0. It is worth noting this inconsistency, but perhaps it does not matter that much to the majority of candidates as they are aiming for a higher score anyway.

On the other hand, some universities that accept both IELTS and PTE for admission purposes ask for higher PTE grades than those Pearson lists as equivalent to IELTS scores. Table 13 is an example of entry requirements of IELTS and PTE scores for international students to study financial mathematics and statistics in five Russell Group universities in the UK.

Table 13: *Example of IELTS and PTE entry requirements by a competitive UG program*

University program	IELTS	PTE
Imperial College Mathematics with Statistics for Finance	7.0 (all four bands 6.5)	69 (all four skills 62)
LSE Financial Mathematics and Statistics	7.0 (all four bands 7.0)	69 (all four skills 69)
University College London Statistics, Economics and Finance	7.0 (all four bands 6.5)	69 (all four skills 62)
University of Warwick Mathematics, Operational Research, Statistics and Economics (MORSE)	6.5 (all four bands 6.0)	69 (all four skills 59)
University of Bristol Mathematics with Statistics for Finance	6.5 (all four bands 6.0)	67 (all four skills 60)

Source: Websites of these universities (correct for 2020 admissions cycle)

Taking the PTE concordance table as a reference point, we can see these universities require 0.25 to 0.75 IELTS band points higher, which in effect means 0.5 to 1.0 band points higher because IELTS Overall Band score is rounded up. From our analysis of the features of the two tests in terms of their content, constructs, and measurement characteristics/conditions, we agree with the judgement of these five universities with regard to the score equivalence of the two tests.

6. Discussions and conclusion

Kolen and Brennan (2014) suggested that one way to think about linking any two tests is 'in terms of degrees of similarity' in test features, because 'the utility and reasonableness of any linking depends upon the degree to which tests share common features' (p.498). They recommended analysis of at least four key features of tests: populations, constructs, measurement characteristics or conditions, and inferences.

Who are the target populations of IELTS and PTE? Are the two tests measuring the same constructs? To what extent do IELTS and PTE share similar construct in each of the four skills – Listening, Reading, Speaking, and Writing? To what extent are the two tests measuring each of the four skills in a way that reflects their targeted language use context? To what extent are the two tests using similar assessment tasks to measure each of the four skills? What inferences do test-takers and other score users (e.g. governments and universities) make about test scores? To what extent do test-takers and other score users agree with the official recommendations made by test providers?

Weir's (2005) socio-cognitive framework was used to compare the four key features of the two tests to answer these questions. Multiple sources of data were analysed, including official sample tasks, interviews with three candidates who have taken both tests, research articles and promotional publications by and/or on IELTS and PTE.

We know that decisions on English language proficiency requirements are not necessarily always based on research evidence, because there are many other competing factors that institutions take into consideration when deciding their English language proficiency requirements. It is also true that many factors influence their decisions on whether to accept the equivalence table on IELTS and PTE provided by Pearson (Table 12). Their interpretations of equivalence of test results between IELTS and PTE reflect the value, views, and understandings they have about the two tests.

Test-takers, however, have only one aim; they are only concerned about whether they can meet the requirements set by the institutions to which they are submitting their test scores. As a result, the actual test-taker population of the two tests is very much determined by the extent to which the potential test-takers find or perceive which test is easier to achieve the scores they need for their purposes. It is also affected by the extent to which test-score users (e.g. universities and governments) are implementing the Pearson recommended equivalence table or with further institutional adjustment or discretion. Even with the higher PTE scores required, test-takers may still find it easier to meet the PTE than IELTS requirements, which may further increase the competition for test-takers between the two tests.

Test providers have the responsibility to educate score users (especially governments and universities) with solid research evidence to facilitate comparison of the two tests, through rigorous and independent studies. However, such studies should go much beyond to simply produce an equivalence or concordance table or any correlation statistics.

In fact, such concordance tables are problematic in many aspects. Firstly, they assume that IELTS and PTE have shared constructs of assessment to a good extent, which may not be the case as our analysis demonstrated. Secondly, any statistical modelling on the relationships between the two test scores assumes that there is sufficient transparency of the scoring systems and procedures of both tests. As our analysis demonstrated, PTE does not provide full details on the weightings of each task and each enabling skills that contribute to the calculation of the four communicative/language skill scores. Thirdly, this kind of correlational study can shed light on only one aspect of the comparability of the two tests, that is, the end-product of test performance. Fourthly, the statistical analysis of test performance as groups, for example, in relation to bands/scores, first language, age and gender, can make the differences between individuals less noticeable.

The richness of the interview data collected from the three candidates who have taken both IELTS and PTE multiple times can confirm the enormous variations at individual levels in their test-taking and preparation strategies, processes, and attitudes between candidates. Our interview data clearly showed there are big differences between them in their English language proficiency, purpose of taking the tests, and their test preparation processes, however, it is possible that they may be considered belonging to the same group, e.g., in relation to their first language, in a statistical analysis. Studies on candidates' different preparation (see Yu & Green, 2021) and test-taking strategies and processes (e.g. whether they had previously taken the tests, how many times they repeated the tests, how they prepared for the tests, and why they were taking the tests), from the viewpoint of candidates themselves, would equally, or perhaps even more importantly, demonstrate the level of comparability between the two tests.

With the correlational study as a starting point, more robust empirical studies embracing both quantitative and qualitative data (e.g. interviews with test-takers of both tests, see Appendix 1: Interviews with IELTS and PTE test-takers) would help us better understand the degrees of similarity between the two tests from the perspectives of individual test-takers. However, it should be noted, as we argued above, that an equivalence table itself, especially if it uses only the overall score/band, is not sufficient to capture the big differences in the constructs of assessment and the operationalisation of the constructs between PTE and IELTS.

It should also be noted that an equivalence table can over-generalise and make the big differences between the two tests invisible to score users. The correlational study by the IELTS Partners (Elliot & Blackhurst, 2021) has already shown clearly that the correlations in the four language skills and at different score band/level between the two tests varied to quite a high degree, even though the correlation in overall score/band between the two tests was found to be similar to what was reported by PTE earlier. Based on data from the field test of PTE and participants' self-report scores of IELTS (n=2432), Zheng and de Jong (2011) reported a moderately high correlation between PTE and IELTS candidates' overall performance (r=0.76 based on self-report scores, and r=0.73 based on data who provided official score reports, n=169). Both correlational studies suggested that around 50% variance in the participants' PTE performance (in the overall score) could be explained by their IELTS performance (overall band). However, given the use of very different assessment methods by the two tests, a more fine-tuned equivalence table that includes not only the equivalence in overall score/band and four language components, but also at different band/score level, is much needed, to reflect the big differences in constructs and measurement characteristics between the two tests.
It is also desirable to make a much fine-tuned and detailed comparison of, e.g. test-takers' performance at a question type as well as a set of question types in a skill (Speaking, Writing, Listening, and Reading), in order to address questions like 'to what extent do candidates' performance in the two tests differ in the same and different question types?'.

The differences between the two tests are noticeable in many aspects. Although both IELTS and PTE assess the four language skills (Speaking, Writing, Listening, and Reading) the constructs of these skills and methods of assessment differ to a large extent. The two tests place different degrees of emphasis on different aspects of the four language skills and use very different methods to assess them. They are structured differently, have different measurement characteristics, and use different task types, scoring methods and criteria. There are only a small number of task/question types that both tests use; the majority of question types used in the two tests are unique to their respective test.

Although we are only able to make broad-brush comparisons between the question types in the two tests, the findings of such comparisons do demonstrate the big differences within a PTE paper in terms of the assessment focus and linguistic and cognitive demands at a level of question/task type. At surface level, some PTE tasks (e.g. summarize written text, summarize spoken text, retell lecture, describe image) do look more academic, authentic and demanding, however, the challenges from these tasks can probably be cancelled out by (a) those easier tasks which assess candidates' performance only at a local and lexical level (e.g. read aloud, repeat sentence, dictation of a sentence or single words) and (b) the possibility that the more demanding tasks might have a lower weighting. Wei and Zheng's (2017) finding that the best predictors of the Listening score of 5,000 PTE candidates are the easiest Listening tasks rather than the more challenging ones such as 'summarize spoken text' and 'retell lecture' raised a series of questions on how much each task type in a paper, and how much each language skill in integrated assessment tasks contribute to the assessment of the six enabling skills and the four communicative/language skills (see Section 5.3: Inferences).

Another major difference between IELTS and PTE is in relation to the level of information provided on scoring methods and weighting of each task and task type. IELTS is transparent in how each question is marked, how each question contributes to the assessment of any language skill, and how the overall band is calculated. However, less information is available for PTE. The *PTE Score Guide* (version 12) only provides brief information on how each task is assigned a score. However, candidates are given a different number of tasks to complete. For example, one candidate may be asked to complete two 'write essay' tasks and one 'summarize written text' task and two 'summarize spoken text' tasks, while another candidate may be given one 'write essay' task and three 'summarize written text' tasks and one 'summarize spoken text' task.

During the process of textual analysis of each assessment task of PTE, we were not able to find information on the weighting and parameters that the automated scoring engines use. Candidates are not informed how exactly the overall score is calculated, nor are they informed how exactly each sub-score of the enabling skills contributes to the overall score and the scores of the four language skills (but see PTE-funded research which supports this practice of reporting sub-scores that are fairly highly intercorrelated (Reckase & Xu, 2015)). Echoing the recommendation by Jin and Zhang (2014), one of the first few Pearson-funded research projects on PTE-Academic, we argue that official information on the weighting of each task and each component of the integrated tasks which are designed to assess more than one language skill should be made available to candidates and other score users.

> "It is suggested that the target (or primary) modality/skill of each integrated task be explicitly stated and appropriately weighed if the current practice of reporting the four communicative skills is to be retained. For example, the task read aloud is targeted more specifically at speaking. The reported score of this task should give more weight to speaking than reading. Similarly, score report of the task summarize written text should give more weight to writing than reading. Decisions on weighting are, nonetheless, difficult to make, and would perhaps be largely based on experience.

> However, given the increasing popularity of integrated tasks in language testing and assessment, further exploration of the way to report performance on integrated tasks will prove a worthy effort." (Jin and Zhang, 2014, p.14)

In conclusion, it is evident that the two tests serve similar populations and purposes and have some commonalities in the constructs of the four language skills, however, the operationalisation of the constructs varied to a large extent. Several assessment methods are unique to PTE. Integrated assessment is a prominent feature of several PTE tasks (e.g. summarize written text, summarize spoken text, retell lecture, and describe image), which are also linguistically and cognitively more demanding than other tasks. The difficulty level of IELTS tasks is more balanced across the papers. The difficulty level of the PTE tasks, however, varies to a greater extent. Overall, the cognitive and linguistic demands of the two tests are broadly at a similar level. The lack of information on weighting and the automated scoring engines used by PTE makes it difficult to conduct equating exercises meaningfully. However, as a starting point, the textual analyses have helped us better understand the degrees of similarity so that any future equating exercises can be more targeted.

The findings from the textual analyses also support a more fine-tuned equivalence table(s) which should incorporate not only the overall scores/bands, but also the four language skills separately, at different band/score level, and even at a question type and a set of question types. In addition, the equating exercises should include more qualitative data from test-takers, teachers of test preparation courses, and test score users.

References

Barkaoui, K. (2019). Examining Sources of Variability in Repeaters' L2 Writing Scores: The Case of the PTE Academic Writing Section, *Language Testing*, 36(1), 3–25. https://doi.org/10.1177/0265532217750692

Elliot, M., & Blackhurst, A. (2021). *Investigating the Relationship between Pearson PTE Scores and IELTS Bands.* Cambridge: Cambridge Assessment English. Accessed from: https://www.ielts.org/-/media/research-reports/ielts-pte-comparisons.ashx

Geranpayeh, A., & Taylor, L. (2013). *Examining Listening: Research and Practice in Assessing Second Language Listening*. Cambridge: Cambridge University Press.

Huhta, A., & Randell, E. (1996). Multiple-Choice Summary: A Measure of Text Comprehension. In A. Cumming & R. Berwick (Eds), *Validation in Language Testing* (pp.94–110), Multilingual Matters.

Jin, Y., & Zhang, X. (2014). *Effects of Skill Integration on Language Assessment: A Comparative Study of Pearson Test of English Academic and Internet-based College English Test Band-6*.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge University Press.

Knoch, U., Huisman, A., Elder, C., Kong, X., & McKenna, A. (2020). Drawing on Repeat Test Takers to Study Test Preparation Practices and Their Links to Score Gains, *Language Testing*. https://doi.org/10.1177/0265532220927407

Kolen, M. J. (2007). Data Collection Designs and Linking Procedures. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds), *Linking and Aligning Scores and Scales* (31–55), Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer Science & Business Media.

McCray, G., & Brunfaut, T. (2018). Investigating the Construct Measured by Banked Gap-Fill Items: Evidence from Eye-Tracking, *Language Testing*, 35(1), 51–73. https://doi.org/10.1177/0265532216677105

Nakatsuhara, F., Inoue, C., Lam, D., & Khabbazbashi, N. (2018). *Towards New Avenues for the IELTS Speaking Test: Insights from a Comprehensive Literature Review*. University of Bedfordshire.

Pearson. (2019a). *Pearson Test of English Academic: Automated Scoring.* https://pearsonpte.com/wp-content/uploads/2018/06/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf

Pearson. (2019b). *PTE Academic Score Guide for Test Takers*. https://pearsonpte.com/wp-content/uploads/2019/10/Score-Guide-for-test-takers-V12-20191030.pdf

Reckase, M. D., & Xu, J-R. (2015). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners, *Educational and Psychological Measurement*, 75(5), 805–825. https://doi.org/10.1177/0013164414554416

Rukthong, A., & Brunfaut, T. (2020). Is Anybody Listening? The Nature of Second Language Listening in Integrated Listening-to-Summarize Tasks, *Language Testing*, 37(1), 31–53. https://doi.org/10.1177/0265532219871470

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge University Press.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A Comparison of Three- and Four-Option English Tests for University Entrance Selection Purposes in Japan, *Language Testing*, 23(1), 35–57.

Taylor, L. (2011). *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Cambridge University Press.

Taylor, L. (2013). *Testing Reading through Summary: Investigating Summary Completion Tasks for Assessing Reading Comprehension Ability*. Cambridge University Press.

Taylor, L., & Chan, S. (2015). *IELTS Equivalence Research Project (Gmc 133)*. https://www.gmc-uk.org/-/media/documents/GMC_Final_Report___Main_report__extended____Final___13May2015.pdf_63506590.pdf

Taylor, L., & Weir, C. J. (2012). *IELTS Collected Papers 2: Research in Reading and Listening Assessment*. Cambridge University Press.

van Moere, A. (2012). A Psycholinguistic Approach to Oral Language Assessment, *Language Testing*, 29(3), 325–344. https://doi.org/10.1177/0265532211424478

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. https://doi.org/10.1177/0265532212448619

Wei, W., & Zheng, Y. (2017). An Investigation of Integrative and Independent Listening Test Tasks in a Computerised Academic English Test, *Computer Assisted Language Learning*, 30(8), 864–883. https://doi.org/10.1080/09588221.2017.1373131

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Palgrave Macmillan.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Examinations, 1913–2012*. Cambridge University Press.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited Imitation as a Measure of Second Language Proficiency: A Narrative Review and Meta-Analysis, *Language Testing*, 33(4), 497–528. https://doi.org/10.1177/0265532215594643

Yu, G. (2013a). From Integrative to Integrated Language Assessment: Are We There Yet?, *Language Assessment Quarterly*, 10(1), 110–114. https://doi.org/10.1080/15434303.2013.766744

Yu, G. (2013b). The Use of Summarization Tasks: Some Lexical and Conceptual Analyses, *Language Assessment Quarterly*, 10(1), 96–109. https://doi.org/10.1080/15434303.2012.750659

Yu, G. & Green, A. (2021). Preparing for admissions tests in English, *Assessment in Education*, 28(1) 1–12.

Yu, G., He, L., & Isaacs, T. (2017). The Cognitive Processes of Taking IELTS Academic Writing Task 1: An Eye-Tracking Study. *IELTS Research Reports Online Series*, No.2. IELTS Partners: British Council, IDP: IELTS Australia and Cambridge English Language Assessment. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-2.ashx

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). Preparing for the Speaking Tasks of the TOEFL iBT® Test: An Investigation of the Journeys of Chinese Test Takers, *ETS Research Report Series*, 2017(1), 1–59.

Yu, G., Rea-Dickins, P. M., & Kiely, R. (2011). The Cognitive Processes of Taking IELTS Academic Writing Task One. *IELTS Research Reports,* Volume 11 (373–449), IDP: IELTS Australia & British Council.

Zheng, Y., & de Jong, J. H. A. L. (2011). Establishing Construct and Concurrent Validity of Pearson Test of English Academic. https://pearsonpte.com/wp-content/uploads/2014/07/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf

Appendix 1: Interviews with IELTS and PTE test-takers

1. When did you take IELTS and PTE, and why? Which test did you take first? How many times did you take IELTS and PTE? What were the test results? Do you know any friends/colleagues who have also taken both IELTS and PTE? Do you know their reasons for taking both tests?

2. Could you reflect on your experience of preparing for IELTS and PTE? How many months did you spend before taking IELTS and PTE the first time? Overall, which paper/component (Listening, Reading, Writing, and Speaking) of IELTS did you find most challenging to prepare for? Why? Overall, which paper/component (Listening, Reading, Speaking and Writing) of PTE did you find most challenging to prepare for? Why? Did you attend any test preparation course?

3. Could you reflect on your experience of taking IELTS and PTE? Overall, which paper/component (Listening, Reading, Writing, and Speaking) of IELTS and PTE did you find most challenging? Why?

4. What do you think are the similarities and differences between IELTS and PTE in their Listening test? Which test, IELTS or PTE, do you find more challenging? Why? Which task(s) in IELTS Listening test do you find most challenging? Why? Which task(s) in PTE Listening test do you find most challenging? Why?

5. What do you think are the similarities and differences between IELTS and PTE in their Reading test? Which test, IELTS or PTE, do you find more challenging? Why? Which task(s) in IELTS Reading test do you find most challenging? Why? Which task(s) in PTE Reading test do you find most challenging? Why?

6. What do you think are the similarities and differences between IELTS and PTE in their Writing test? Which test, IELTS or PTE, do you find more challenging? Why? Which task(s) in IELTS Writing test do you find most challenging? Why? Which task(s) in PTE Writing test do you find most challenging? Why?

7. What do you think are the similarities and differences between IELTS and PTE in their Speaking test? Which test, IELTS or PTE, do you find more challenging? Why? Which task(s) in IELTS Speaking test do you find most challenging? Why? Which task(s) in PTE Speaking test do you find most challenging? Why?

8. What do you think IELTS and PTE can learn from each other, for example, in test design and delivery? What would be your advice to them?

REPORT 2

Aligning IELTS and PTE-Academic: A Measurement Study

Mark Elliot, Andy Blackhurst, Barry O'Sullivan, Tony Clark, Jamie Dunlea & Nick Saville

How to cite this study:
Elliot, M., Blackhurst, A., O'Sullivan, B., Clark, T., Dunlea, J., & Saville, N. (2021). Aligning IELTS and PTE-Academic: A measurement study. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series,* No. 2, (pp. 42–63). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Abstract

The research study reported in this paper is designed to be read alongside that of Yu (2021). This is because the two studies present different aspects of the alignment process, Yu taking a qualitative approach, focusing on the underlying construct, while the current study takes a quantitative approach, focusing more on the measurement qualities of both tests.

In this study, we examine the score data from 523 candidates who had taken both tests. The primary statistical process used was equipercentile linking as this approach has the merit of allowing differences in difficulty to vary along the score scale, that is to say with equipercentile equating one test form could be relatively more difficult at high and low scores, but relatively less difficult at the middle scores (Kolen & Brennan, 2014).

Findings suggest that while there appears to be a reasonably stable linear correlation across the two tests in terms of the overall scores reported, these actually mask some extremely different profiles when we look to the sub-scores for each of the four skills. While there are major differences with the productive skills, it is in the area of writing that the most serious issues arise. The lack of available information about the PTE-A scoring system means that we do not know how the overall scores are calculated. While this makes it quite difficult to interpret the relationship between the two tests fully, we present what we believe is a useful alignment table which includes an estimation of the relationship across the four skills as well as the overall scores awarded.

Putative *alignment of IELTS bands and Pearson PTE scores*

IELTS	PTE-A overall	IELTS & PTE-A Listening	IELTS & PTE-A Reading	IELTS & PTE-A Speaking	IELTS & PTE-A Writing
5	40.8	40.2	43	40.2	43.1
5.5	45.4	42.7	47.9	42.2	51
6	51.6	48.1	53.5	46.2	62.2
6.5	58.5	56.8	60.6	53.5	74.1
7	66.3	66.2	67.8	65.3	82.3
7.5	74.6	73.9	73.7	75.3	87.5
8	82.3	79.4	78.4	80.9	89.4
8.5	88.1	84.7	83.7	85.5	89.5

Authors' biodata

Mark Elliot

Mark Elliot is a Senior Assessment Quality and Validity Manager. Since joining Cambridge Assessment English in 2008, he has been involved in a broad range of assessment and validation activities, including test construction and operational data analysis. He has an MA in English Language Teaching and Applied Linguistics from Kings College London and is currently a PhD candidate at Cambridge University, where he also completed an undergraduate degree.

Andy Blackhurst

Dr Andrew Blackhurst is a Principal Research and Validation Manager at Cambridge Assessment English, coordinating test validation activities since 2003. Regular duties include monitoring of live IELTS test performance data and materials, in addition to managing key operational research matters across the IELTS network. His PhD is from University College London, completed in 2001.

Barry O'Sullivan

Professor Barry O'Sullivan is the British Council's Head of Assessment Research & Development. He has worked on numerous test development and validation projects and advises ministries and institutions on assessment policy and practice. His work includes the development and refinement of the socio-cognitive model of test development and validation, and the design and development of the Aptis test. He has presented at many conferences around the world and has over 100 publications. He is the founding president of the UK Association of Language Testing and Assessment, holds a visiting professorship at the University of Reading, UK, and is Advisory Professor at Shanghai Jiao Tong University, China. He has been a fellow of the Academy of Social Sciences in the UK since 2016, and the Asian Association for Language Assessment since 2017. In 2019 he was awarded an OBE by the Government of the UK for his contribution to language testing.

Tony Clark

Dr Tony Clark is a Principal Research Manager at Cambridge Assessment English, focused on the IELTS test. His research interests include language assessment and pedagogy, particularly writing acquisition, standard setting, accommodations, test preparation, diagnostic assessment and lexical studies. These interests largely emerged from pedagogical experience. He has published in and reviews for major testing journals, and holds a PhD from Bristol University. Since joining Cambridge English, he has been responsible for the IELTS Joint-funded Research Program and the Caroline Clapham Master's Award, acted as Permanent Secretary/Chair of the IELTS Joint Research Committee, and led on a number of high profile cross-partner research projects, in addition to several standard-setting workshops. Tony's PhD received a British Council Research Assessment Award, and he was a recipient of both Newton Fund and ESRC Scholarships. Prior to joining Cambridge English, he contributed to research projects on admissions testing, language acquisition and test development.

Jamie Dunlea

Dr Jamie Dunlea is a senior researcher and manager of the British Council's Assessment Research Group. He has a PhD in language testing from the Centre for Research in English Language Learning and Assessment and works on a range of language test development and validation projects for the British Council, as well as collaborating with researchers and organisations internationally. Jamie has advised Ministries of Education and national agencies on assessment reform projects, overseen research for collaborative, international projects such as linking UK examinations to China's Standards of English (CSE), and is active in the language assessment research community. He joined the British Council in 2013 after heading validation research at the Eiken Foundation, a not-for-profit organisation which develops and administers EFL examinations in Japan. He has over 25 years of experience working in EFL education, first as a teacher, then in test development and assessment research.

Nick Saville

Dr Nick Saville is Director of Research & Thought Leadership at Cambridge Assessment English (University of Cambridge), and is the Secretary-General of the Association of Language Testers in Europe (ALTE). He regularly presents at international conferences and publishes on issues related to language assessment. His research interests include assessment and learning in the digital age; the use of ethical AI; language policy and multilingualism; the CEFR; and Learning Oriented Assessment. Nick was a founding associate editor of the journal, *Language Assessment Quarterly*. He is joint editor of *Studies in Language Testing* (SiLT, CUP) and editor of the *English Profile Studies* series (EPS, CUP). He is a member of the IELTS Research and Development Steering Group, and sits on several University of Cambridge Boards, including: the Interdisciplinary Research Centre for Language Sciences; the Institute for Automated Language Teaching and Assessment (ALTA); and English Language iTutoring (ELiT), providing AI-informed automated systems.

Contents

1. Introduction 46

2. Aligning tests 46

2.1 Quantitative-only studies 47

2.2 Qualitative and quantitative studies 48

3. The current study 49

4. Methodology 49

4.1 Participants 49

5. Analysis 50

6. Results 51

6.1 Scatterplots 51

6.2 Equipercentile graphs 53

6.3 Comparing the current study with Clesham & Hughes (2020) 58

6.4 An alternative alignment table 59

7. Conclusions 60

7.1 Interpreting results across concordance tables 60

7.2 Integrating quantitative and qualitative data: Summarising the results of the current study and Yu (2021) 61

7.3 Limitations 61

References 62

List of tables

Table 1: Distribution of sample by IELTS band score 50

Table 2: Concordance table for overall scores 53

Table 3: Concordance table for Listening 54

Table 4: Concordance table for Reading 55

Table 5: Concordance table for Speaking 55

Table 6: Concordance table for Writing 56

Table 7: Correlations between scores on IELTS and scores on PTE 58

Table 8: Putative alignment of IELTS bands and Pearson PTE scores, based on Clesham & Hughes (2020: 11) 58

Table 9: Putative alignment of IELTS bands and Pearson PTE scores 60

List of figures

Figure 1: Scatterplot of overall scores in IELTS and PTE-A 51

Figure 2: Scatterplot of Listening scores in IELTS and PTE-A 51

Figure 3: Scatterplot of Reading scores in IELTS and PTE-A 52

Figure 4: Scatterplot of Speaking scores in IELTS and PTE-A 52

Figure 5: Scatterplot of Writing scores in IELTS and PTE-A 53

Figure 6: Equipercentile graph for overall scores 54

Figure 7 Equipercentile graph for Listening 54

Figure 8: Equipercentile graph for Reading 55

Figure 9: Equipercentile graph for Speaking 56

Figure 10: Equipercentile graph for Writing 57

Figure 11: Equipercentile graph for overall plus four skills 57

Figure 12: Graphical representation of the alignment claims 59

1. Introduction

Over the past number of years, the IELTS Academic test has been aligned successfully to the Common European Framework of Reference for Languages (CEFR) – see, for example, Lim et al (2013) – and China's Scale of English (CSE) – see Dunlea et al (2019). Building the CEFR into the test development and revision processes is critical to establishing a stable link between the two; see Hawkey & Barker (2004) and O'Sullivan (2015).

In recent decades, test score users (governments, higher education institutions, employers, etc.) have looked to internationally-recognised tests to inform their decisions around such things as migration, university entrance (and exit) and recruitment. Until relatively recently, the range of options open to test users was small, and the decision as to which tests to accept was relatively easy – users typically opted for the familiar, so in the US the default test was usually seen as the Test of English as a Foreign Language (TOEFL), while in the UK and Europe the default tests were IELTS or Cambridge English examinations such as C1 Advanced and C2 Proficiency.

A decade ago, O'Sullivan (2011: 7) described changes in the world of assessment in terms of professionalisation (with many MA and PhD students completing their studies and returning home with the knowledge and skills obtained) and localisation (the appreciation that the test-taker lies at the centre of the test and an increased awareness of the importance of context). He argued that these changes were leading to a fragmentation of the language testing industry. In the intervening decade, this fragmentation has continued apace.

A significant issue that has arisen with the emergence of new tests is that users are now faced with a bewildering number of comparisons when deciding on the most appropriate test or tests for their situation. There is then an urgent need for test users to be provided with evidence of the alignment of different tests. This urgency is highlighted by the proliferation of overly simplistic concordance tables without sufficient empirical support, making decisions very difficult for users to make. Too often, the concordance table is based only on a quite simplistic comparison of the scores awarded and fails to take into consideration the actual content of the tests. We therefore find that tests that are in no way comparable are claimed to be aligned, often based on correlation data.

In response to this situation, the IELTS Partners proposed a two-part approach to alignment: one focuses on the test content and the degree to which it reflects a particular language model, and the other focuses on the measurement qualities of the tests (what we will refer to as the construct study). The Partners decided at that same time to undertake a comparison between IELTS and a relative newcomer to the international language testing market, the Pearson Test of English Academic (PTE-A), taking this approach (this is the measurement study). The construct study was undertaken by Professor Guoxing Yu of the University of Bristol, while the measurement study was undertaken by the validation team at Cambridge Assessment English. These two studies have been summarised by Elliot & Blackhurst (2021).

2. Aligning tests

This review offers a brief overview of a range of comparative test analysis and criterion-related validation studies which are relevant to the current study. This is not meant to be a comprehensive literature review, but instead is designed to showcase the range of studies that have been undertaken over the past number of decades. We present a selection of studies based on the quality of the work and the approach taken – from the most basic (small number of participants and simplistic analysis) to the more complex, integrating qualitative and quantitative approaches with reasonable populations.

2.1 Quantitative-only studies

In 2015, the developers of the Duolingo English Test (DET) published a study which included what they described as "preliminary linking results" between the DET and IELTS and by extension, the CEFR and TOEFL (Bézy & Settles, 2015: 3). The claims were based on what appears to be an extremely small set of data. The paper (op. cit., 2) states that "36 submitted IELTS scores, 1 submitted TOEFL scores, and 2 submitted both". Despite this, correlation analysis was undertaken, and tables included claiming alignment to the CEFR, and comparing the DET to both IELTS and TOEFL. No later study has been published by the developers. Reviews of the DET (Kunnan & Wagner, 2015; Wagner, 2020) suggest that the lack of any meaningful model of academic language makes the test unsuitable for use for university admissions. This also suggests that a critical review of the underlying constructs of the three tests (and the CEFR) would have ruled out any meaningful comparison study before it could be undertaken.

The Language Testing & Training Center (LTTC) in Taiwan first introduced the General English Proficiency Test (GEPT) 20 years ago. In that time, the research team there have conducted four studies in which they compared GEPT to other English language tests. Their approach changed over the years from the initial quantitative approach taken by Chin and Wu (2001) in their comparison of the GEPT Intermediate and the EIKEN Grade 2 speaking tests. The EIKEN tests were Japan's most widely used English-language testing program and remain so to this day. Their study found similarities in terms of difficulty level across the two tests, despite significant differences in format.

A later study (LTTC, 2003) compared performance (again quantitatively) on the GEPT and two other tests, the TOEFL Computer-based Test (CBT) and the College English Test band 6 (CET-6). Test data analysis indicated medium to high correlations across the tests, though with quite different profiles across the different papers (e.g. while the CBT Listening paper was found to be more difficult than that of both the High-Intermediate GEPT and CET-6, when it came to the Reading paper, the CBT was found to be the easiest, and CET-6 the hardest). These findings act as a warning to developers who carry out comparison studies that the overall test score (an amalgam of the scores on the constituent papers) can hide important differences at the individual paper or skill level.

Weir et al (2013) examined what they referred to as the criterion-related validity, in line with Weir's 2005 socio-cognitive frameworks, of the Reading and Writing papers from the Advanced level GEPT when compared with the same components of IELTS and also with later actual academic performance. While the latter type of comparison can be criticised due to the complexity of identifying the role of language in overall academic performance, the results of the test comparison reflected the findings of the LTTC (2003) study in concluding that it was more difficult to achieve the equivalence of a CEFR Level C1 on the GEPT papers.

Another important study in this regard is that of Brown et al (2012) who examined the alignment of the EIKEN tests mentioned above and the TOEFL Internet-based Test (iBT) using a range of quantitative techniques. These techniques included a Rasch model for equating and linking and correlation analyses and Principal Components Analysis to explore similarities around the underlying constructs of the two tests. In this way, the study represents a more sophisticated, though still purely quantitative approach.

2.2 Qualitative and quantitative studies

Perhaps the most well-known comparison study to have been undertaken in the area of language testing is the Cambridge – TOEFL Comparability Study (Bachman, Davidson, Ryan & Choi, 1995). The focus of the study was to draw comparisons based on qualitative and quantitative analyses of the Cambridge First Certificate in English (FCE) (now known as B2 First) and the Test of English as a Foreign Language (TOEFL) administered by the Educational Testing Service (ETS) – note that the study pre-dated the introduction of the TOEFL Internet-based Test (iBT) so involved the paper-based version of the test. Despite some obvious issues, such as the comparison of a general English proficiency test (FCE) and a test of English for Academic Purposes (TOEFL) with little reference to this difference, the combination of the qualitative, construct-focused element with the quantitative, measurement-focused was to set a standard for future studies that has rarely been met in the intervening years.

Although limited to the Reading paper, Wu (2014) employed a more comprehensive integration of qualitative and quantitative methods in a comparison of GEPT Reading tests at CEFR B1 and B2 levels with Cambridge tests targeting the same levels. As reported in Wu et al (2016), "The results indicated that the Intermediate GEPT and Preliminary English Test (PET), both of which target B1 level, were in general comparable, while the High-Intermediate GEPT and First Certificate in English (FCE), which target B2 level, exhibited greater differences, not only in terms of test results, but also in contextual features and cognitive processing operations."

Wu et al (2016) took this mixed methods approach when attempting to draw comparisons between the GEPT and the British Council's Aptis. The two tests are different in target level, with Aptis a single instrument that tests across multiple proficiency levels, while the GEPT offers a suite of level-specific tests. For this reason, the population for this study comprised 144 candidates across four GEPT tests (Elementary, Intermediate, High-intermediate, and Advanced). Results suggest that there are relatively high correlations across all subtests.

Dunlea et al (2018) describe another major comparability study in which a two-pronged approach was taken. Here the comparison was made between the Aptis and the VSTEP, which is recognised by universities in Vietnam as certification of English proficiency for the purpose of meeting graduation requirements stipulated by the Ministry of Education in Vietnam. The VSTEP targets CEFR levels B1 to C1 (Dunlea et al, 2019, p.7). The approach taken in this study was based on the socio-cognitive model (Weir, 2005; O'Sullivan & Weir, 2011) and consisted of sophisticated statistical analyses (including Rasch and factor analysis), together with a comprehensive evaluation of the underlying constructs using a specially designed framework and questionnaire data from candidates.

In late 2020, Pearson published a study which included an updated concordance table for PTE-A and IELTS (Clesham & Hughes, 2020). This table was based on a quantitative study of the scores of 562 candidates on the two tests. Of this population, just over half provided their official IELTS score report, an additional 105 reported their IELTS sub-scores while all overall and sub-scores were known for PTE-A. The resulting table, which suggested some significant changes to the original concordance table (Zheng & De Jong, 2011) will be discussed below, in the Results section. Note that the original concordance table referred to by Clesham & Hughes (2020, 11) was not actually included in the earlier report though Zheng & De Jong (2011, p.36) indicated that "concordance coefficients were generated between PTE Academic and other tests of English using linear regression" based on the overall scores/bands achieved by their participants. They went on to state that these "regression coefficients were then used to predict the scores of PTE Academic BETA test-takers' scores on TOEFL iBT and IELTS" before claiming that "two complete concordance tables have been generated based on the established conversion coefficients, one among PTE Academic, TOEFL iBT scores, and CEF, the other among PTE Academic, IELTS, and CEF."

Since the two studies referred to above were focused, at least in part, on PTE-A and IELTS, we will return to discuss both in relation to the findings of this study in the Conclusions section below.

3. The current study

This paper is part of a two-part project in which separate research teams explored the relationship between the IELTS and the Pearson Test of English – Academic (PTE-A) through different methodology. The first part in this volume (pp. 7–41) takes the form of a comprehensive qualitative construct study of the two tests (Yu, 2021) and complements this study. This paper reports on the quantitative study undertaken as part of the project. We will discuss the Yu study in the Conclusions section of this paper.

4. Methodology

As indicated above, this paper takes a quantitative approach to the alignment of the two tests. It is expected that readers will read it alongside the Yu (2021) qualitative study to gain a more fully balanced overview of the alignment and to fully interpret the summary of the findings from the two studies contained in the Conclusions section below.

4.1 Participants

This project grew from a survey of test-taker experiences with different tests undertaken for IDP: IELTS Australia by Catalyst Research, an independent research firm based in Perth, Australia, working with Macquarie University International College English Language Centre, during which participants who had taken both IELTS and PTE-A within 90 days were asked to provide score information. Given the interest in the relationship between scores on the two tests, it was decided to extend the quantitative dimension and Catalyst was engaged to expand the data sample.

Reflecting the initial (and continuing) focus on Australia, the largest cohort of participants (377) within the final sample took their IELTS test there. While this did provide a diverse sample in its own right (35 nationalities were represented in this Australian sample alone), further participants were recruited currently, who had taken the test elsewhere, most notably the UAE (49 participants) and India (34 participants), together with other participants who had taken their IELTS test in China/Hong Kong, Nepal, Pakistan, the United Kingdom and United States.

In total, score information was obtained from 523 test-takers who had taken both IELTS and PTE-A within 90 days of each other. However, not all participants provided complete sets of sub-scores for the four skills, so analysis was based on 519 individuals at overall score level; 404 for Listening; 404 for Reading; 405 for Writing; and 404 for Speaking. As noted, the sample came from a suitably diverse range of nationalities and first languages. In one respect, however, the recruitment of participants was unsatisfactory. It was never intended that the sample should reflect the ability distribution of the wider IELTS candidature as the project design envisaged approximately equal numbers of participants at each of the bands 5 to 8. In the event, however, recruitment was much more successful among higher performing test-takers. The sample distribution by IELTS band score is given in Table 1.

Table 1: *Distribution of sample by IELTS band score*

Ability distribution of participants	
Overall band score	**% of test-takers**
<5	0.8%
5	2.1%
5.5	6.2%
6	12.7%
6.5	14.0%
7	19.0%
7.5	23.1%
8	17.7%
>8	4.4%

5. Analysis

The equipercentile linking method was employed to compare results on the two tests, following the model established in the IELTS/Cambridge English Qualifications comparisons reported in Lim et al (2013) and paralleling the equipercentile linking method employed in the Pearson PTE/IELTS comparison study reported in Clesham & Hughes (2020). As discussed in Kolen & Brennan (2014), the equipercentile approach has the merit of allowing differences in difficulty to vary along the score scale, that is to say with equipercentile equating one test form could be relatively more difficult at high and low scores, but relatively less difficult at the middle scores. Equivalence is established by identifying scores on one test that have the same percentile ranks as on the other, such that for any given score on one test the percentage of test-takers securing that score or a lower score is established and then the score (or lower) on the other test secured by the same percentage of test-takers is identified. These two scores are then deemed to be equivalent, as representing the same standard of achievement.

Analysis was carried out using RAGE-RGEQUATE (Zeng et al, 2004), errors estimated using the Equating Error software (Hanson & Chien, 2004), and appropriate models selected for each of the four skills. To counter the possibility of distortions which might arise from relatively small and therefore not necessarily completely representative samples (Kolen & Brennan 2004), smoothing methods have been developed to produce estimates of the distributions and equipercentile relationships having the smoothness property that would characterise the broader test-taking population. As in the IELTS/Cambridge English Qualifications project, it was decided to use pre-smoothing and to utilise the polynomial log-linear method, available within RAGE-RGEQUATE, which fits polynomial functions to the log of sample density (Holland & Thayer, 2000). This method of smoothing was adopted because indices are available for evaluating goodness of fit and appropriateness of the linking (Kolen & Brennan, 2004).

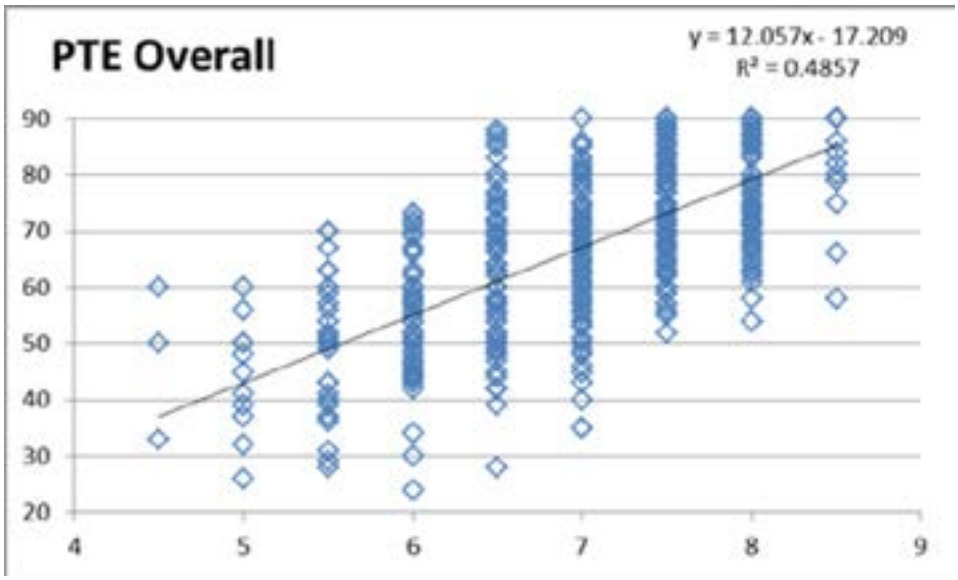
6. Results

In this section we present the results of the analyses undertaken. These will be presented first as a series of scatterplots that are designed to offer a broad picture of the relationship across the scores awarded on the two tests. We will then further explore the data through the lens of the equipercentile graphs, which again offer an interesting perspective on the data from the two tests. Finally, we turn to the concordance table, focusing on the similarities and significant differences across the two tests.

6.1 Scatterplots

The scatterplot for the overall scores on the two tests can be found in Figure 1. Here, we can see that there is a medium-strength relationship between the two tests (R^2=0.4857) though there are relatively few data points for the lower score range, i.e. below 5.5.

Figure 1: *Scatterplot of overall scores in IELTS and PTE-A*



When we turn to the scatterplots (Figures 2 to 5) for the reported scores for the four skills (Listening, Reading, Speaking and Writing), we can see that these are again positive though quite different in profile.

Figure 2: *Scatterplot of Listening scores in IELTS and PTE-A*

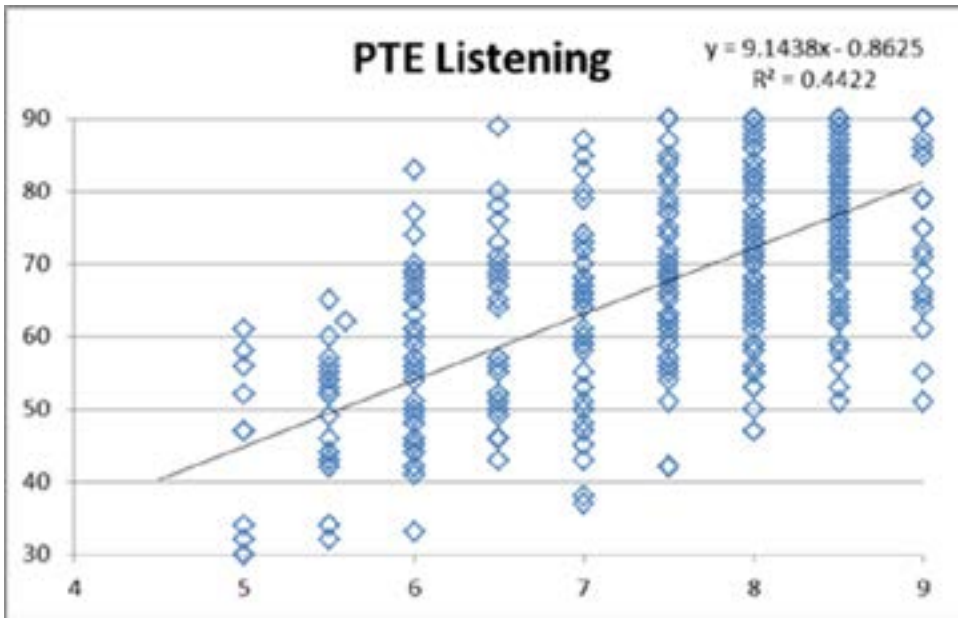
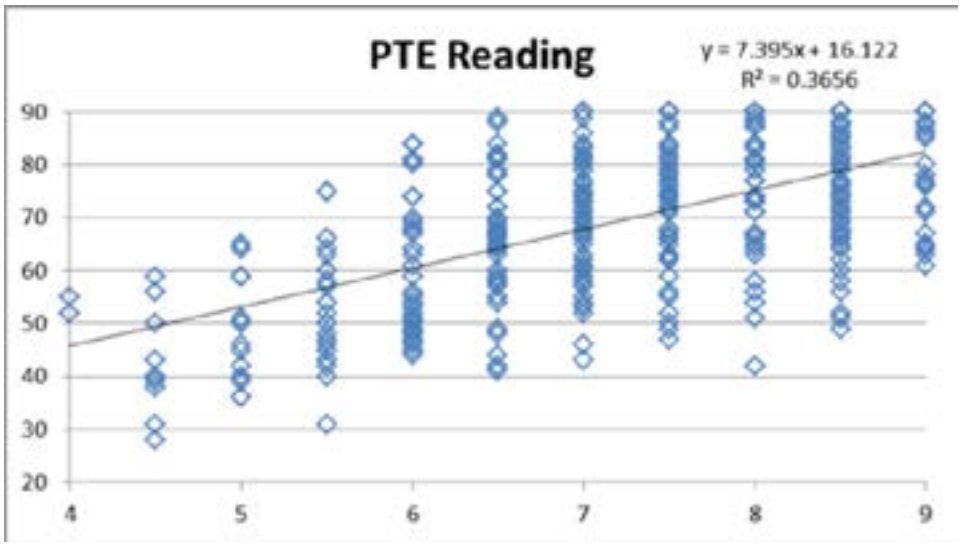
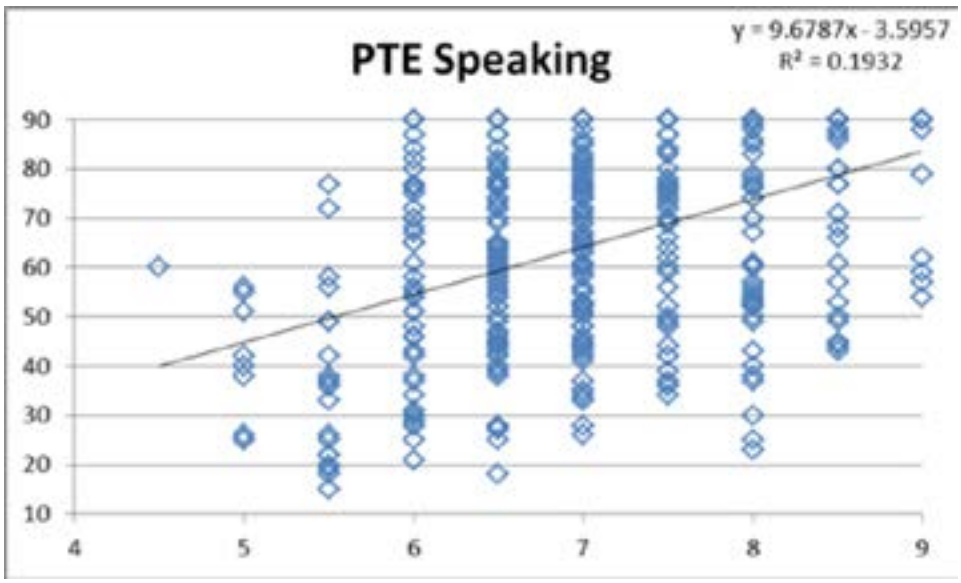


Figure 3: *Scatterplot of Reading scores in IELTS and PTE-A*



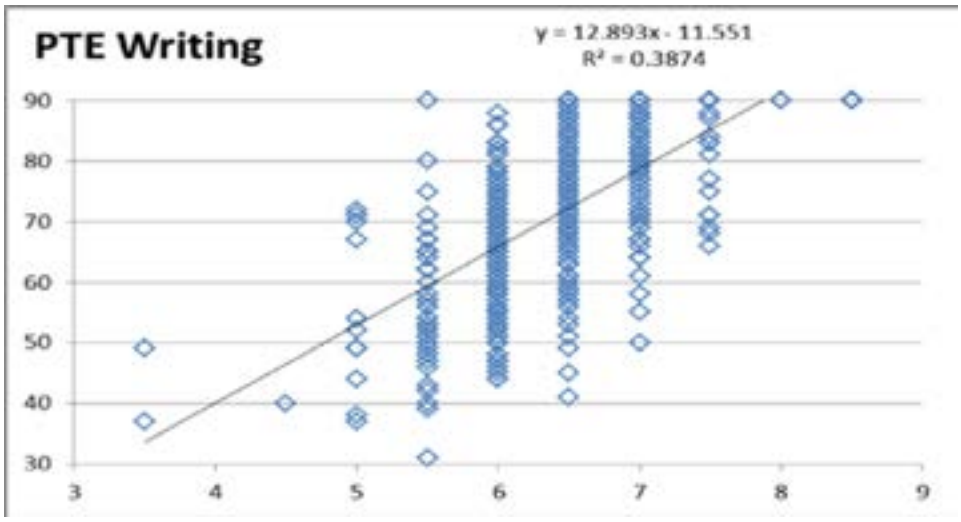
Perhaps not surprisingly, neither of the receptive skills reflect the same R^2 value as the overall, since overall is an amalgam of the four scores we would always expect that it would be higher that the individual component scores. The trend lines, however, suggest that there are different patterns of performance across the two tests, again unsurprising given that they are quite different in focus and format (see Yu, 2021).

Figure 4: *Scatterplot of Speaking scores in IELTS and PTE-A*



It is when we get to the productive skills that we find the biggest differences. Figure 4 indicates that the Speaking scores have a very low R^2 estimate, indicating a positive though low correlation between the scores on the two tests. This suggests that comparison of test performances on this skill may be problematic. Given the issues raised by Yu (2021) in this regard, the indication is that test users may need to be cautious when drawing comparisons for Speaking across the tests.

Figure 5: *Scatterplot of Writing scores in IELTS and PTE-A*



Equipercentile looks at the distribution of scores right across the scale, so while the R^2 value for Writing (Figure 5) is not low (it is actually higher than that for Reading), the range appears to be significantly truncated at the higher end, with the PTE-A Writing effectively topping out at the IELTS band 8 level. This suggests that there is a significant issue here. We will return to this below.

6.2 Equipercentile graphs

Equipercentile graphs offer a useful visual representation of the estimated relationship across the two tests. We present a series of concordance tables (Tables 2–5) and related graphic representations (Figures 6–10) each of which offer a valuable perspective on the relationship between the two tests based on the overall scores and on the four skills.

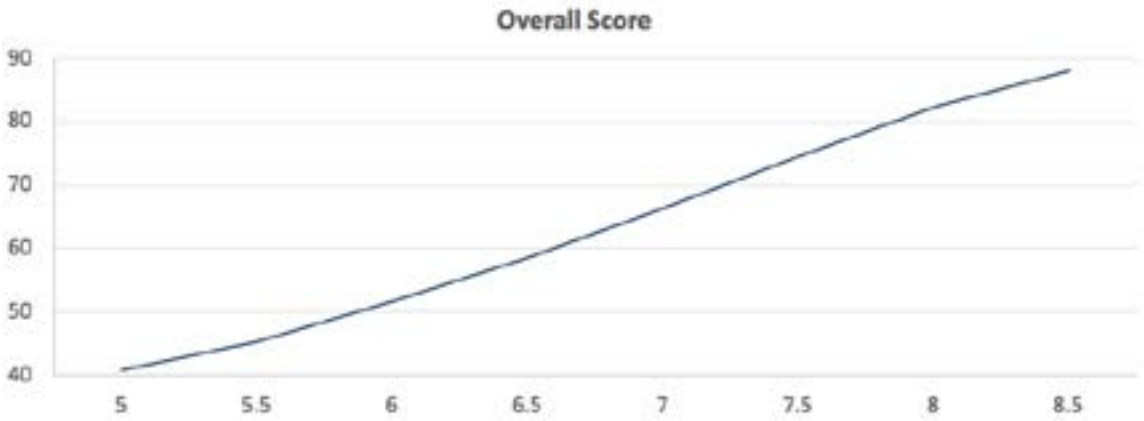
6.2.1 Overall Score

The concordance table for the overall score (Table 2) shows a relatively even set of steps across the two tables, suggesting a somewhat linear relationship. This relationship is confirmed in the chart that follows (Figure 6) and suggests that the rationale offered by Yu (2021) in support of an alignment argument is confirmed. In order to further explore the relationship between the two tests, it is necessary to review the findings for the four skills as reported by the IELTS Partnership and Pearson.

Table 2: *Concordance table for overall scores*

	Scale	yx	se	se.b
1	5.0	40.76146	0.8628093	0.7845668
2	5.5	45.35398	1.2665579	1.0969325
3	6.0	51.58694	1.2318055	1.1678655
4	6.5	58.53999	1.2932963	1.2280964
5	7.0	66.27297	1.2587240	1.1635235
6	7.5	74.55021	1.0962364	0.9920956
7	8.0	82.30825	0.9493683	0.8584286
8	8.5	88.11916	0.7309762	0.6512508

Figure 6: *Equipercentile graph for overall scores*



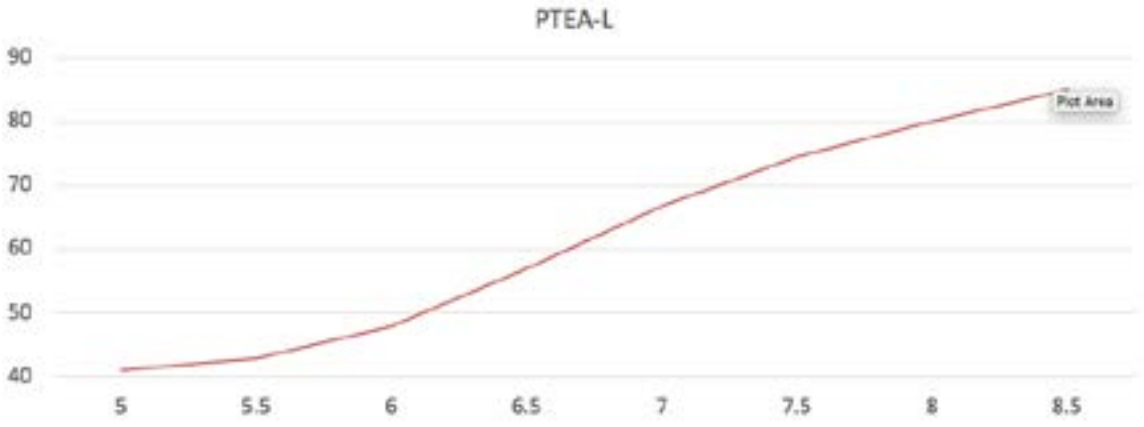
6.2.2 Listening

The concordance table for Listening (Table 3) suggests that the relationship between the two tests is less clear than the overall scores suggest. There is a clear bottoming out effect to be found for the PTE-A Listening paper below the IELTS band 6 level. While the relationship is somewhat linear above this level, there would appear to be a question mark around the use of the PTE-A Listening scores for decisions below 5.5 or 6. The addition of data at the lower levels would clarify this situation. As expected, the chart (Figure 7) confirms this finding.

Table 3: *Concordance table for Listening*

	Scale	yx	se	se.b
1	5.0	40.23707	0.7229789	0.6332728
2	5.5	42.71233	1.2388990	1.1330317
3	6.0	48.12857	1.5669816	1.4317567
4	6.5	56.75870	1.6074360	1.4533589
5	7.0	66.24173	1.5107784	1.3718328
6	7.5	73.94971	1.3876561	1.2961874
7	8.0	79.43488	1.3223454	1.2885461
8	8.5	84.73362	1.4078582	1.1861980

Figure 7: *Equipercentile graph for Listening*



6.2.3 Reading

The concordance table for Reading (Table 4) indicates that this paper demonstrates the closest relationship across the four skills. The steps from IELTS level to level are all relatively equal – in fact they range from approximately 5–8 points on the PTE scale. This relationship is confirmed in the related chart (Figure 8) with its representation of an almost linear relationship.

Table 4: *Concordance table for Reading*

	Scale	yx	se	se.b
1	5.0	42.99891	1.410367	1.2067798
2	5.5	47.89908	1.501075	1.4593999
3	6.0	53.49646	1.655884	1.5682711
4	6.5	60.55533	1.666954	1.6147487
5	7.0	67.84451	1.491879	1.4686746
6	7.5	73.73299	1.252517	1.2428936
7	8.0	78.35382	1.096181	1.0768625
8	8.5	83.69480	1.090230	0.9061987

Figure 8: *Equipercentile graph for Reading*

6.2.4 Speaking

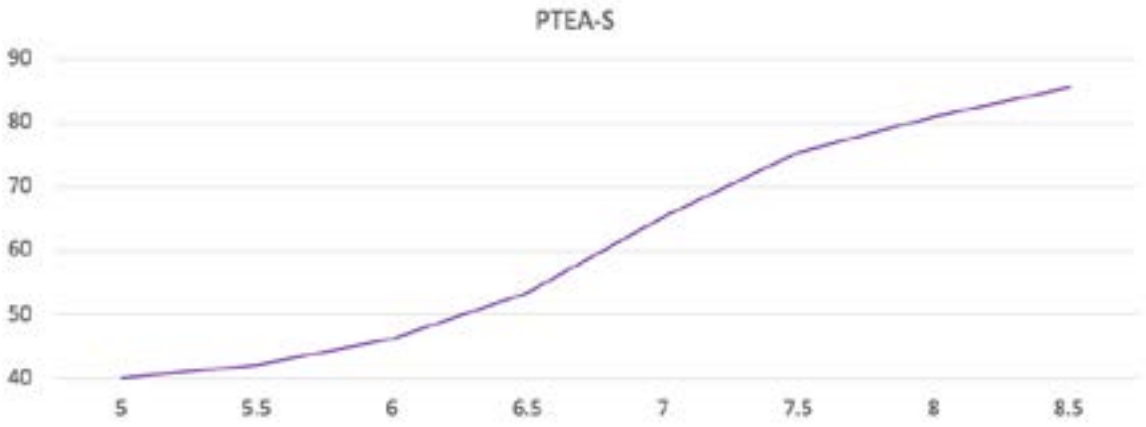
The concordance table for Speaking (Table 5) reflects to a large extent what is happening with the Listening data. Here again, we see that the relationship is clearly curvilinear, in fact almost an S-shape. The data indicates that the relationship between the two Speaking papers is not easily interpreted from a measurement perspective.

Table 5: *Concordance table for Speaking*

	Scale	yx	se	se.b
1	5.0	40.15496	0.6589449	0.5723585
2	5.5	42.17077	1.0516333	0.9467973
3	6.0	46.20474	1.2438123	1.1304001
4	6.5	53.46676	1.6919827	1.6524944
5	7.0	65.25109	2.2849466	2.1773222
6	7.5	75.32197	1.6737388	1.6209173
7	8.0	80.90768	1.3292715	1.2748341
8	8.5	85.50931	1.3106047	1.0289886

The findings from our analysis of Table 5, are highlighted in Figure 9, where we can clearly see that the two tests can really only be considered for mutual interpretability between IELTS band 6.5 and 7.5. It appears that the PTE-A Speaking paper again awards higher-than-expected scores at the lower levels, perhaps due to the task types described by Yu (2021) which allow lower-level candidates to gain points in their scoring system. Since we do not know how the overall scores for the four skills are estimated, we cannot be certain of why this finding occurs in the data.

Figure 9: *Equipercentile graph for Speaking*



6.2.5 Writing

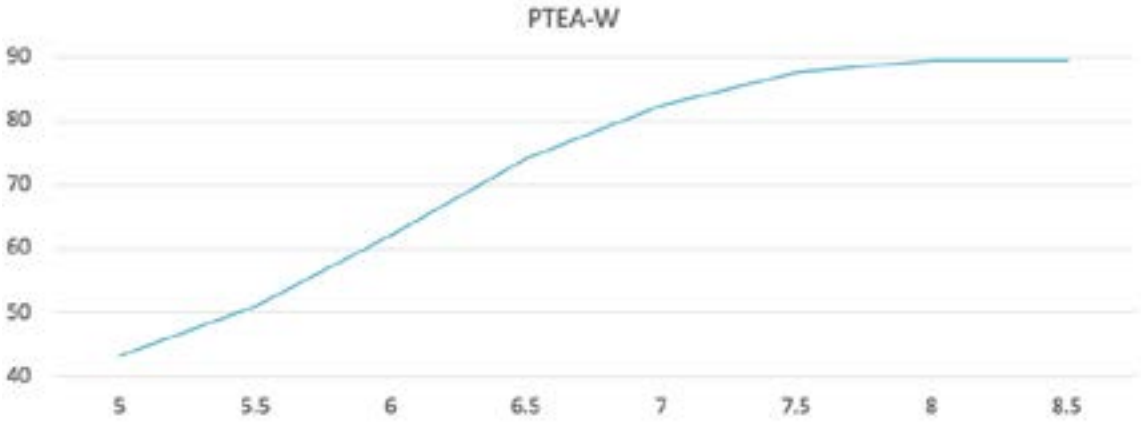
The concordance table for Writing (Table 6) suggests that there is a relatively linear relationship between the two tests up to the level of IELTS band 6.0, though the rise in PTE scores appears to be at a greater rate than seen with the other skills. After that, the PTE scores taper off until there is little or no movement score-wise in relation to IELTS. This is because there appears to be a significant topping-off effect for the PTE-A scores.

Table 6: *Concordance table for Writing*

	Scale	yx	se	se.b
1	5.0	43.13323	1.71946867	1.39411761
2	5.5	50.97456	1.38290617	1.20679319
3	6.0	62.15329	1.49294727	1.32356210
4	6.5	74.06259	1.18313320	1.07792821
5	7.0	82.32697	0.93661065	0.87295713
6	7.5	87.50599	0.97416557	0.70536971
7	8.0	89.36798	0.22983379	0.29954861
8	8.5	89.49843	0.02360109	0.01744959

This finding is highlighted in the chart (Figure 10). Here it is obvious that the PTE-A Writing scores are topping out by IELTS band 8. In fact, given that the SEM reported for Pearson is at an overall level of 2.3 GSE points, and that the SEM for Writing and Speaking are always lower than for the receptive skills, it appears that candidates are likely to achieve a full score (90 points) on the PTE-A Writing paper for a score as low as 7.5 on the IELTS paper. We do not have an SEM estimate for the PTE-A skills scores, but this finding indicates that there is a significant issue with the way the Writing paper is scored.

Figure 10: *Equipercentile graph for Writing*

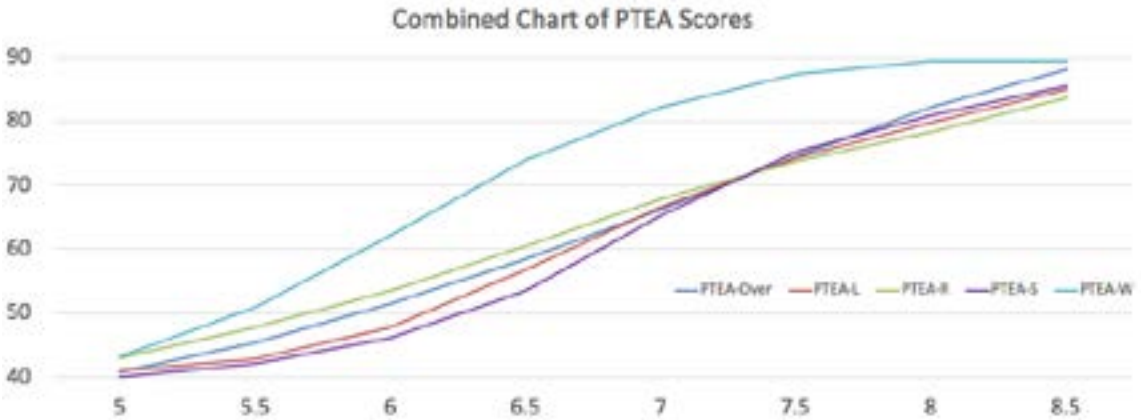


6.2.5 Overview

When we put the equipercentile graphs together, we can see the extent of the problems with linking the two tests. The Overall estimates hide the fact that there are clear differences to be found between the scores awarded for the productive skills, though the receptive skills are close in terms of profile. This is perhaps less problematic for the Speaking paper, particularly as from approximately IELTS band 7, the relationship across the two tests appears stable in terms of scores awarded. Essentially, the graphs show us that a relatively low level of gain in terms of PTE-A Speaking score can result in a significant move up the IELTS scale. This is exemplified by the fact that a move from approximately 40 to 42 on the PTE-A scale (lower than the SEM reported for the test as a whole) sees a jump from 5 to 5.5 on the IELTS scale while another 4 points on the PTE scale will take the candidate to a 6.

However, the Writing paper clearly stands apart as the problematic paper. The scoring profile will be of real concern for those attempting to interpret what the Writing paper is actually testing, and the interpretation of the scores awarded on that paper. There is clearly something different impacting on the IELTS–PTE-A relationship for productive skills as the profile is so radically different.

Figure 11: *Equipercentile graph for overall plus four skills*



6.3 Comparing the current study with Clesham & Hughes (2020)

Since the current study and that of Clesham & Hughes (2020) are both focused on establishing evidence of an alignment in relation to test scores between IELTS and the PTE-A, we now take some time to compare the findings and draw some conclusions.

We begin this comparison by looking to the correlations reported here and by Clesham & Hughes (2020). The figures reported are remarkably similar, suggesting that we are dealing with two quite similar datasets (Table 7).

Table 7: *Correlations between scores on IELTS and scores on PTE*

Component	Pearson Correlation This study	Pearson Correlation Clesham & Hughes (2020)
Overall	0.70	0.74
Listening	0.66	0.66
Reading	0.60	0.68
Speaking	0.44	0.42
Writing	0.62	0.60

We next turn to the alignment claims from the Clesham & Hughes (2020) report and this study. Table 8 takes the reported concordance from Clesham & Hughes (2020) and adds a column, PTE (this study), to allow for a comparison of claims. It is clear from a comparison of the PTE (updated) column from Clesham & Hughes (2020) that there are some similarities around the IELTS 6.5 decision point. It is equally clear, however, that there are significant differences from this point down.

Table 8: *Putative alignment of IELTS bands and Pearson PTE scores, based on Clesham & Hughes (2020: 11)*

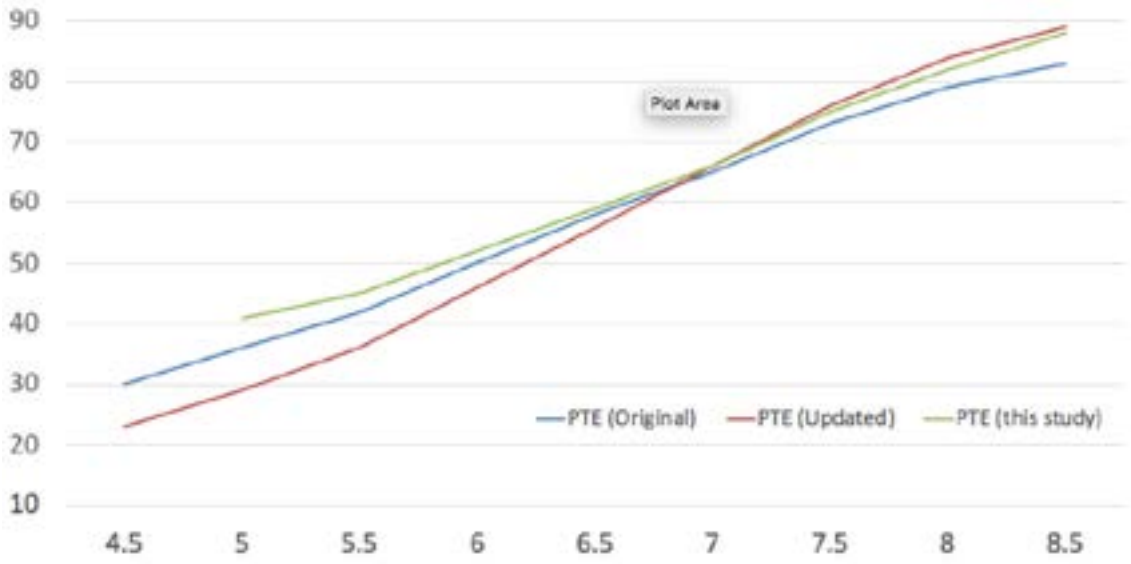
IELTS	PTE (original)	PTE (updated)	PTE (this study)
4.5	30	23	
5.0	36	29	41
5.5	42	36	45
6.0	50	46	52
6.5	58	56	59
7.0	65	66	66
7.5	73	76	75
8.0	79	84	82
8.5	83	89	88

The graphical representation of this table highlights this latter issue (Figure 12). It is obvious from this representation that the updated equivalences from Pearson are significantly lower than originally estimated. Note that for an IELTS band 6 (one of the important cut-scores for university entrance and migration decisions) the change from the original is 4 points on the Pearson Global Scale. This is quite a lot higher than the SEM for the PTE-A, reported by Clesham & Hughes (2020) as 2.3 points on the scale, and as such represents a significant, and unexplained, change.

The difference between the estimate from this study and the updated PTE study is 6 points. This implies that a person applying for a university place or a working visa requires a full half band lower from PTE-A than from IELTS. Where institutions or ministries accept lower proficiency levels (e.g. 5 for admission to preparatory programs), the situation is even more problematic. We estimate that the difference here is 12 points on the Pearson Scale, or approximately 3.5 to 4.0 on the IELTS scale.

At the other end of the scale, while there are changes evident in the data in Table 6 and the Chart in Figure 12 around the key decision points (i.e. 6.5 to 7.5), there is again a significant shift in the relationship reported in Clesham & Hughes (2020) from IELTS grade 8 up. It is clear that the Clesham & Hughes (2020) report heightens the requirement in a meaningful way at the 8.0 and 8.5 levels in particular to the extent that the updated requirement appears to almost match the estimation from the current study. Again, while they are clearly moving in the right direction, these changes are significantly greater than the SEM for the test and require further attention.

Figure 12: *Graphical representation of the alignment claims*



6.4 An alternative alignment table

It is quite clear from the above tables and charts that a comparison of the overall scores from two tests is likely to result in some level of confusion with regard to the true alignment between the tests. We saw that very similar correlation outputs, which would lead many to imply a strong link between the tests, hides a number of significant issues. This phenomenon was also reported by Yu (2021) who suggests that similar uses and populations, together with a broadly similar assessment approach suggest that it is appropriate to continue with an alignment project. However, Yu (2021) later presents evidence that demonstrates the many differences between the two tests (as well as a number of similarities, of course). We will return to Yu later in this paper to consider the results of his work in combination with the current study.

Given the evidence above, we suggest that a more detailed alignment table should be presented in order to allow test users to view the detailed evidence they require, especially at the policy level. This is included here as Table 9.

The interpretation of this table is straightforward. The first two columns on the left present the alignment of the tests at the overall score level. To interpret the other columns, identify the skill you are interested in, identify the IELTS level, then look across to the appropriate skill column – so for example the alignment of the tests of Reading at IELTS 6.5 is 60.6 on the Pearson scale. Another example would be to look at IELTS 6.5 Writing, which equals PTE 74.1 for Writing, and PTE Overall at 58.5.

Table 9: *Putative alignment of IELTS bands and Pearson PTE scores*

IELTS	PTE-A overall	IELTS & PTE-A Listening	IELTS & PTE-A Reading	IELTS & PTE-A Speaking	IELTS & PTE-A Writing
5	40.8	40.2	43	40.2	43.1
5.5	45.4	42.7	47.9	42.2	51
6	51.6	48.1	53.5	46.2	62.2
6.5	58.5	56.8	60.6	53.5	74.1
7	66.3	66.2	67.8	65.3	82.3
7.5	74.6	73.9	73.7	75.3	87.5
8	82.3	79.4	78.4	80.9	89.4
8.5	88.1	84.7	83.7	85.5	89.5

7. Conclusions

In his detailed qualitative comparison of the underlying constructs of the two tests, Yu (2021) found that there appears to be a difference in difficulty across the two tests. Participants in his study reported that the PTE-A is less cognitively and linguistically challenging than IELTS, and the complex score profiles of the participants were identified. The findings reported here suggest that, in terms of reported scores, the complex relationships between IELTS and PTE-A scores can be confirmed.

7.1 Interpreting results across concordance tables

While the comparison of overall scores on both tests suggests that there is a relatively stable and linear relationship between them, additional analyses revealed a number of interesting, and in one case disturbing, issues. These can be summarised as follows.

- Around the 6.5 to 7.5 area, there are some differences, though these tend to lie within the SEM of the PTE-A so are unlikely to be of significant concern.
- Below the IELTS 6.5 level, the difference appears to grow as the level decreases. This suggests that test users should review their current policies where decisions are made regarding migration and study below this proficiency level. The differences between results of Clesham & Hughes (2020) and this study range from 0.5 to 1.0 IELTS bands.
- Above the IELTS 7.5 level, the difference has actually been narrowed to the extent that there is little difference across the two tests.
- Significant changes, in whichever direction, from one alignment table to the next (particularly where these are above one SEM) should be very carefully explained so that test score users can be confident that scores they accept will not have adverse consequences for their systems.

7.2 Integrating quantitative and qualitative data: Summarising the results of the current study and Yu (2021)

Despite the caveats that have been pointed out in this report, the multi-method approach taken here allows us to draw a number of conclusions from the studies. The most obvious of these are as follows.

- The data indicate that the PTE Writing paper is significantly different at the upper end of the reporting scale than the IELTS Writing paper. The 'topping out' effect seen in Figure 10 shows that the typical PTE candidate will reach the top of the scoring scale when at the IELTS Band 7.5 level.
- This issue impacts on the overall scores awarded for what would be similar levels of performance on the two tests compared here.
- There appears to be a significant difference in the way in which the Speaking skill is tested and scored across the two tests. The tendency of the Speaking test to result in quite different profiles is indicated by the correlation coefficients presented both by IELTS and Pearson (see Table 5). The suggestion here is that test score users should carefully review the two Speaking tests to identify the most appropriate for their context.

While the concordance table presented here (Table 2) tells us a lot about the relationship between the two tests, the qualitative data offers a vital additional insight. The IELTS Partnership therefore recommends that test users refer to the table when making decisions, but at the same time, we believe that it is necessary to look beyond the numbers to understand more fully the strengths and weaknesses of tests that are presented to them.

7.3 Limitations

As with any research study, there are a number of limitations in the current work, some related to the approach taken and others to the quality of the data and information available. These can be summarised as:

- The population is very similar in size and quality to that of the other studies reported here. However, as with these studies, the sample tends to be self-selecting to some extent and while it is broadly representative of the test population this cannot be fully established in reality.

Descriptions of the samples for both this study and Clesham & Hughes (2020) highlight that the number of participants drops noticeably at lower levels. Future studies should try to achieve a more balanced sample across proficiency levels (while recognising that this is difficult to achieve in practice).

References

Bachman, L. F., Davidson, F., Ryan, K. & Choi, I-C. (1995). An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge TOEFL Comparability Study, *Studies in Language Testing: Vol. 1*. Cambridge: Cambridge University Press.

Bézy, M. & Settles, B. (2015). The Duolingo English Test and East Africa: Preliminary linking results with IELTS & CEFR, *Duolingo Research Report DRR-15-01*. Accessed from: https://s3.amazonaws.com/duolingo-papers/reports/DRR-15-01.pdf

Brown, J. D., Davis, J. McE., Takahashi, C. & Nakamura, K. (2012). *Upper-level EIKEN Examinations: Linking, Validating and Predicting TOEFL iBT Scores at Advanced Proficiency EIKEN Levels.* Society for Testing English Proficiency, Tokyo, Japan. Accessed from: https://www.eiken.or.jp/eiken/group/result/pdf/eiken-toeflibt-report.pdf

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in *PMLR*, 81:77–91.

Chin, J. & Wu, J. (2001). STEP and GEPT: A concurrent study of Taiwanese EFL learners' performance on two tests. *Proceedings of the Fourth International Conference on English Language Testing in Asia*, 22–44.

Clesham, R. & Hughes, S. R. (2020). *2020 Concordance Report: PTE Academic and IELTS Academic*. London: Pearson. Accessed from: https://pearsonpte.com/wp-content/uploads/2020/12/2020-concordance-Report-for-research-pages.pdf

Dunlea, J., Spiby, R., Wu, S., Zhang, J. & Cheng, M. (2019). China's Standards of English Language Ability (CSE): Linking UK Exams to the CSE. *Technical Report VS/2019/003*. London: British Council. Assessed from: https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf

Dunlea, J., Spiby, R., Quynh Nguyen, T. N., Yen Nguyen, T. Q., Huu Nguyen, T. M., Thao Nguyen, T. P., Thuy Thai, H. L. & Sao, B. T. (2018). Aptis–VSTEP Comparability Study: Investigating the Usage of Two EFL Tests in the Context of Higher Education in Vietnam, *British Council Validation Series, VS/2018/001*. London: British Council. Accessed from: https://www.britishcouncil.org/sites/default/files/aptis-vstep_study.pdf

Elliot, M. & Blackhurst, A. (2021). *Investigating the Relationship between Pearson PTE Scores and IELTS Bands*. Cambridge: Cambridge Assessment English. Accessed from: https://www.ielts.org/-/media/research-reports/ielts-pte-comparisons.ashx

Hanson, B. A., & Chien, Y. (2004). Equating Error Computer Software. Iowa City: Iowa: CASMA.

Hawkey, R. & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9: 122–159.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions, *Journal of Educational and Behavioral Statistics*, 25, 133–183.

Jo, E. S., and Gebru, T. (2020). *Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In Conference on Fairness, Accountability, and Transparency* (FAT* '20), 27–30 January 2020, Barcelona, Spain. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3351095.3372829

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Language Training and Testing Center. (2003). *Concurrent validity studies of the GEPT Intermediate level, GEPT High-Intermediate level, CBT TOEFL, CET-6, and the English test of the R.O.C. College Entrance Examination*. Taipei: Language Training and Testing Center.

Larkin, L. (2017). 'I was trying to decide what accent to use for my re-test' – Irish engineer who failed Australian visa English fluency test marked by automatic program, *Irish Independent Newspape*r online edition. Assessed from: https://www.independent.ie/irish-news/i-was-trying-to-decide-what-accent-to-use-for-my-re-test-irish-engineer-who-failed-australian-visa-english-fluency-test-marked-by-automatic-program-36015370.html

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard Setting to an International Reference Framework: Implications for Theory and Practice, *International Journal of Testing*, 13:1, 32–49, DOI: 10.1080/15305058.2012.678526

O'Sullivan, B. (2011). Introduction. In B. O'Sullivan (ed.), *Language Testing: Theories and Practices* (pp.1–12). Oxford: Palgrave.

O'Sullivan, B. (2015). Linking the Aptis Reporting Scales to the CEFR, *Technical Report TR/2015/003*. London: British Council. Assessed from: https://www.britishcouncil.org/sites/default/files/tech_003_barry_osullivan_linking_aptis_v4_single_pages_0.pdf

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English Test, *Language Assessment Quarterly*, 12:3, 320-331, DOI: 10.1080/15434303.2015.1061530

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019, *Language Assessment Quarterly,* 17:3, 300-315, DOI: 10.1080/15434303.2020.1771343

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach. Research and Practice in Applied Linguistics* (Ed). Basingstoke: Palgrave.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing theories and practices* (pp. 13–32). Basingstoke: Palgrave Macmillan.

Weir, C., Chan, S. H. C., & Nakatsuhara, F. (2013). Examining the Criterion-Related Validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and Real-Life Academic Performance, *LTTC GEPT Research Reports RG-01*, Taipei: Language Training and Testing Center. Accessed from: https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG01.pdf

Wu, R. Y. F. (2014). *Validating Second Language Reading Examinations: Establishing the Validity of the GEPT Through Alignment with the Common European Framework of Reference*. Cambridge: Cambridge University Press.

Wu, R. Y., Yeh, H., Dunlea, J., & Spiby, R. (2016). Aptis–GEPT Test Comparison Study: Looking at two tests from Multi-Perspectives using the Socio-Cognitive Model, *Technical Report VS/2016/002*. London: British Council. Accessed from: https://www.britishcouncil.org/sites/default/files/aptis-gept_trial.pdf

Yu, G. (2021). IELTS Academic and PTE-Academic: Degrees of Similarity. In N. Saville & B. O'Sullivan (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series, No. 2*, (pp. 7–41). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2004). RAGE-RGEQUATE [Computer software], Iowa City: University of Iowa.

Zheng, Y., & De Jong, J. (2011). *Establishing Construct and Concurrent Validity of Pearson Test of English Academic* [Research Note], London: Pearson. Accessed from: http://pearsonpte.com/wp-content/uploads/2014/07/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pd