Alignment of the Global Scale of English to other scales: the concordance between PTE Academic, IELTS, and TOEFL

April 2017

Dr John H.A.L. de Jong
Dr Veronica Benigno

Pearson

Contents

Executive Summary

The question of how different language proficiency tests relate to each other is key to the theoretical discussion on consequential aspects of test validity because it provides an indication of how test scores are used and understood by stakeholders and in local decision-making by individual academic institutions. The Global Scale of English (henceforth: GSE) is a linear transformation of the logit scale underlying the descriptors developed by North (2000) to describe the proficiency levels of the Common European Framework of Reference for Languages (henceforth: CEFR; Council of Europe, 2001). The scale, ranging from 10 to 90, was first used as the reporting scale for the Pearson Test of English Academic (Pearson, 2010) and validated by aligning it to other international proficiency scales such as IELTS and TOEFL (De Jong, 2009a; Pearson, 2010; Zheng & De Jong, 2011; De Jong & Zheng, 2016). This paper provides a summary of the evidence of the link between PTE Academic and two other major high-stake tests, IELTS and TOEFL, with the purpose of showing how TOEFL and IELTS scores relate to GSE scores.

Background

Many studies have been conducted to investigate whether high-stakes tests such as PTE Academic, TOEFL and IELTS are able to predict students' success in their academic studies. Mixed findings of studies on predictive validity suggest that the relationship between test performance and academic performance is difficult to ascertain (Bayliss & Ingram, 2006; Cho & Bridgeman, 2012; Graham, 1987; Elder, 1993; Woodrow, 2006). A number of difficulties arise when attempting to correlate test scores to academic performance due to the fact that a single measure, in this case English proficiency, is not sufficient to provide an accurate picture of academic performance, a multidimensional and complex reality. Two major difficulties can be identified. Firstly, it is impossible to analyse a fully representative data set of test-takers' ability given that individuals below the required minimum entry score do not enter academic life and are therefore not included in the research sample under investigation. Secondly, academic success is not exclusively determined by English proficiency but depends on a number of other variables such as the students' background, their subject of study, the effect of instruction, their preparation and motivation for academic success, and

the measures and criteria universities use to determine their students' success. Nowadays, many researchers point out that high-stakes tests function as gatekeepers (Spolsky, 1997) and this may imply serious consequences for students if the measurement of their language ability is not accurate and reliable.

The validity argument around PTE Academic and its reporting scale, the Global Scale of English, is supported by evidence of: a) the relation of PTE Academic scores to the levels of the CEFR (Council of Europe, 2001), which is based on independent expert ratings of test items and candidate responses (Pearson, 2010); and b) the relation of PTE Academic with TOEFL and IELTS, which is based on a comparison of students' scores achieved on PTE Academic with IELTS/TOEFL scores obtained within two months of their PTE Academic score (Zheng & De Jong, 2011), as well as on data provided by ETS and Cambridge about the concordance between TOEFL and IELTS (Taylor, 2004; Tannenbaum & Wylie, 2008; Liao, Qu & Morgan, 2010; ETS, 2015).

Alignment of PTE Academic GSE scores to CEFR levels

PTE Academic alignment research adhered to the guidelines provided in the *Manual for relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009), following the stages of familiarization, specification, standardisation and validation. The Global Scale of English, the reporting scale first used for reporting PTE Academic scores, was developed during the test development phase by indexing it against the original logit values of the CEFR. The validity of the alignment of PTE Academic to the CEFR was demonstrated by gathering independent expert ratings of test items and candidate responses (Zheng & De Jong, 2011; De Jong & Zheng, 2016). The two independently derived estimates of the correspondence of PTE Academic scores with CEFR proved to be highly correlated (r=0.99). Two approaches were followed:

• An item-centred approach was used to relate the IRT difficulty values of each item to CEFR levels. Item writers developing items for PTE Academic were instructed to write items operationalising the B1, B2 and C1 levels of the CEFR. After submission of the items, the level assignments suggested by the item writers were reviewed by independent raters; where

discrepancy occurred, a third rating was gathered. For each item, the estimated difficulty on the CEFR scale indicated during the item development and item review stages was compared with the observed difficulty value established during the field test.

- A test-taker centred approach was used to analyse the test-takers' ability. Five items from two speaking and one writing item types were selected. For each response, the ability estimate (on the CEFR) provided by two human raters was compared with the ability estimate based on scored responses of the PTE Academic items.

The alignment of the GSE to the CEFR has been corroborated through the extension of the set of descriptors published in the CEFR framework in 2001. Since 2013 Pearson has worked on substantially increasing the number of descriptors originally provided in the CEFR and on scaling them, replicating the methodology used by North (2000). The GSE Learning Objectives for Adult Learners http://www.english.com/blog/gse-learning-objectives-for-adults were developed to overcome some of the limitations of the CEFR, e.g., the scarcity of descriptors at the lower and higher levels and the uneven distribution of descriptors over the four skills. As a result, the Global Scale of English provides more granular information about language proficiency than is possible with the CEFR. Since the CEFR only contains exemplary descriptors, the new GSE Learning Objectives add detail about the linguistic activities required at each level and thereby increase the accuracy of the measurement scale.

The Global Scale of English alignment with the CEFR (see Figure 1) can only be fully understood if it is supported by information showing what it really means to be at a level. Being at a CEFR level is defined as having a 50% probability of being able to perform all language activities at that given level of proficiency. If this proficiency level is defined as an interval on a scale, e.g., B1 on the CEFR, being at B1 means one is expected to be able to perform 50% of all tasks at B1, or to have 50% chance of being able to perform any task at B1. If this proficiency is defined as a point on a scale, e.g. 40 on the GSE, then one is expected to be able to perform 50% of all tasks which are at 40 on the GSE or to have a 50% chance of being able to perform any task at 40 on the GSE.
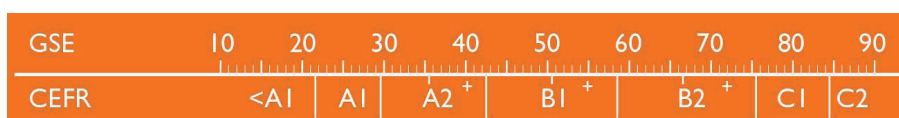
GSE	10	20	30	40	50	60	70	80	90
CEFR		<A1	A1	A2 +	B1 +	B2 +	C1	C2

FIGURE 1

Alignment of the Global Scale of English to the CEFR

Learning a language is not like learning mathematics or electrical engineering, where each topic builds upon a previous one in a logical sequence. Language learning is not necessarily sequential, and a learner might be highly proficient in one area, where they have had extensive practice or where they feel a particular need or motivation, but quite weak in another. For that reason, to say that a learner is 'at' a certain level on the Global Scale of English does not mean he/she has necessarily mastered every GSE Learning Objective for every skill up to that point. Neither does it mean that he/she has mastered none at a higher GSE value. Language learning is unique to every individual and the definition of what it means to be at a level is based on probability: if a learner is considered to be level 61 on the Global Scale of English, he/she has a 50% likelihood of being capable of performing learning objectives at that level – and a greater probability of being able to perform learning objectives at a lower GSE level, such as 50 or 37. As proficiency increases, the probability of being able to perform the learning objectives at 61 also increases (see Figure 2).
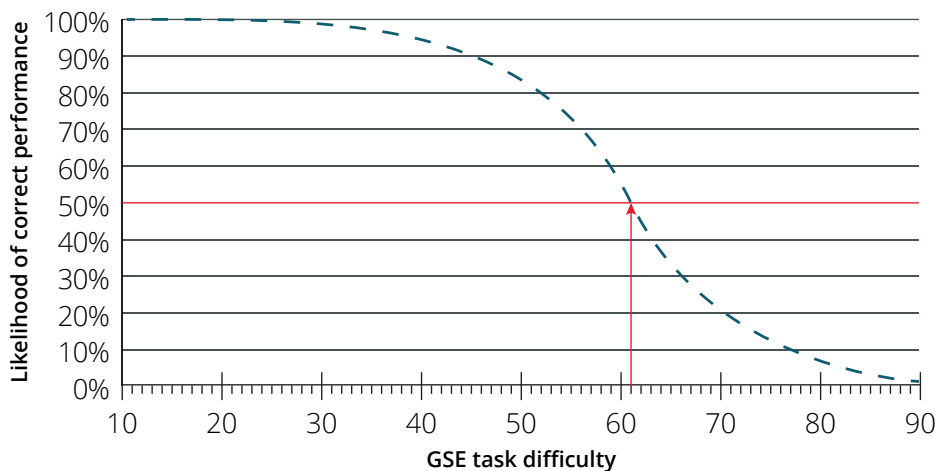

FIGURE 2
A learner at 61 on the GSE

Alignment of PTE Academic GSE scores to IELTS and TOEFL scores

PTE Academic has been field-tested using over 10,000 test-takers. Field testing took place in 2007 and 2008. Test-takers were representative of the global population of students seeking admission to universities and other tertiary education institutions where English is the language of instruction. Test-takers were born in 158 different countries and spoke 126 different languages.

As reported by Zheng and De Jong (2011), a concordance study between PTE Academic, IELTS, and TOEFL was conducted during the field test using test-takers' self-reported scores (accompanied, if available, by a copy of their official score

report). Based on the relationship between scores reported by test-takers and scores obtained in PTE Academic, concordance transformation functions were generated between PTE Academic, TOEFL iBT, and IELTS using best fitting regression functions. This study was repeated during the PTE Academic beta test when again test-takers' scores on the other two tests were collected. The correlation between reported scores and mean scores from the same test-takers as predicted from their PTE Academic score was found to range between .73 and .83 (see Pearson, 2010 for further details).

Additional evidence of the concordance between the three tests was gathered by using data provided by ETS about linking TOEFL iBT to IELTS (ETS, 2015) and to the CEFR (Tannenbaum and Wylie, 2008). Claims about the relation between IELTS and the CEFR are explained on the IELTS website (IELTS, n.d.), but the findings of ETS and De Jong (2009a) suggest the claimed correspondence is at least one CEFR level too high. Data from test-takers were analysed yielding standard descriptive statistics by skill and overall score. Relations among the major English tests and their relations with the CEFR are summarized in the concordance chart presented in the Appendix.

Additional evidence of score reliability

Additional evidence on PTE Academic score reliability and the validity of the GSE can be provided with regard to: a) the test dimensionality; and b) the score reliability as supported by evidence obtained via the analysis of test repeaters' behaviour and in relation to the granularity of the GSE.

PTE Academic dimensionality

PTE Academic validity was first demonstrated in a study conducted by De Jong (2009b) based on field test data and later by a similar study conducted by Reckase & Jing-Ru (2014) on a much larger sample of live test data. De Jong (2009b) analysed the field test data. De Jong showed that although there is essential unidimensionality in a large dataset representative of the total target population, test-takers from different language backgrounds have different profiles. For example, students whose first language is Mandarin tend to score relatively higher on the written skills, whereas test-takers from India tend to score higher on the spoken skills. Reckase and Jing-Ru (2014) corroborated De Jong's findings and suggested that the total score can be reported on a single unidimensional scale while at

the same time reporting subscores provides useful information to help guide instruction/preparation. This second study used data from 37,140 examinees, 954 items, and 164 test forms. The authors found that test-takers with the same L1 showed similar score patterns on specific test items – but different from test-takers with a different L1. This study is important in that it shows that subscores can provide useful information about different constructs for multiple examinee groups even though the test data are considered to be a good fit to a unidimensional model.

Score reliability in comparison to other scales

One piece of evidence of the reliability of PTE Academic scores comes from the analysis of test repeaters' data. Internal investigation of changes in test scores based on the frequency at which the exam is taken recorded no accuracy fluctuation in score when test-takers resit after short periods. A sample of around 2,000 test-takers was analysed showing that 75% of repeaters obtain scores within the 95% error margin on the overall score and 10% test-takers have zero change on repeating PTE Academic. Data are shown in Figures 3a and 3b for TOEFL and PTE Academic respectively. These data are higher than for TOEFL iBT who report data showing zero change for about 5% of test-takers (Zhang, 2008, figure 5, page 10) indicating less score change on PTE Academic than on TOEFL on repeating the test. More detailed information on score changes for test repeaters on PTE Academic are reported by Barkaoui (forthcoming).
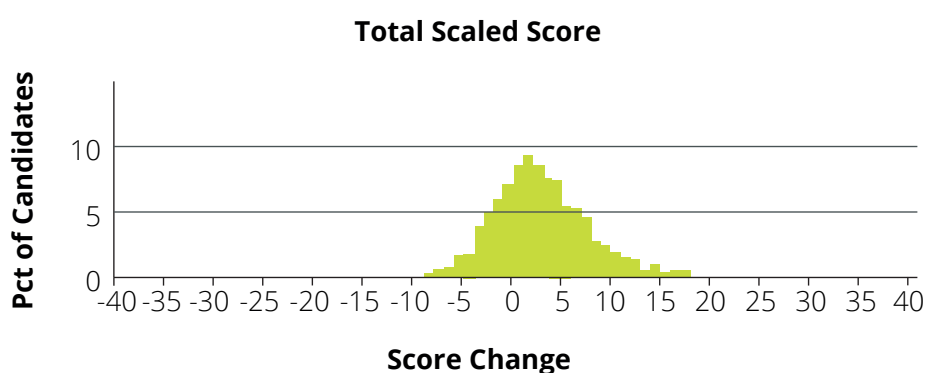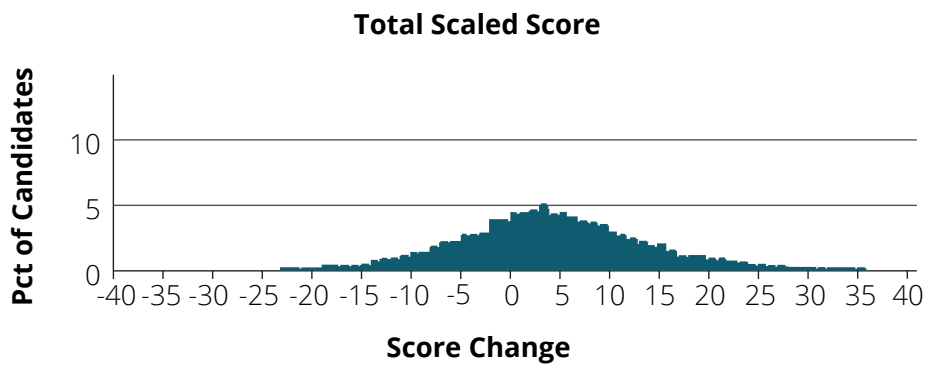
Total Scaled Score

FIGURE 3a

Frequency distributions of total scaled score changes for PTE Academic repeaters

Total Scaled Score

FIGURE 3b

Frequency distributions of total scaled score changes for TOEFL repeaters (Zhang, 2008, figure 5, page 10)

Additionally, internal research has been conducted to compare PTE Academic test design and scoring information with those of TOEFL and IELTS. The relation between PTE Academic and TOEFL iBT scores is corroborated by the agreement of the relation of both instruments with the CEFR. Such corroboration via the CEFR could not be found for the relation between PTE Academic and IELTS scores, although it can be found in the reported score equivalences between TOEFL iBT and IELTS (ETS, 2015). This ETS study can be used as a triangulation: if A=B and B=C, then A=C. In test scores: if, as found in the Pearson concordance study, PTE Academic scores 53-63 correspond to TOEFL iBT scores 79-93, and TOEFL iBT scores 79-93 correspond to IELTS 6.5, then we would expect IELTS 6.5 to correspond to PTE Academic 53-63. Indeed, we found evidence to support this assumption: a study to predict a score of 6.5 on IELTS corresponded to test-takers obtaining 58-64 on PTE Academic.

PTE Academic has been designed to measure all four skills, as both 'pure' skills and integrated with other skills, with around 100 score points per skill and about 250 raw score points for the overall score. The more score points, the greater the accuracy of the measurement. Tables 1 and 2 show the differences between the way speaking and writing abilities are measured in the three tests. PTE Academic presents more points of measurement and therefore shows a greater accuracy. We should also point out that, in terms of test design, IELTS and TOEFL iBT stimuli present scripted language that has been specifically written for the test (no spontaneous live language). PTE Academic uses only 'found' texts, i.e. real texts (written texts or live recordings of speech).

IELTS	TOEFL iBT	PTE Academic
Three task types (independent); one task from each and variable number of questions; total of 3 tasks	Three task types (one independent, two integrated); two tasks from each; total of 6 tasks	Five task types (one independent, four integrated); between 3 and 10 tasks each; total of 35 tasks
Max raw score 4; 'criterion areas' rated on 9 band scales: 4 x 9 = 36; arithmetically reduced to a band of 0 to 9 speaking score; raw score divided by 4 = reported band score	Max raw score 6; tasks each rated on 4 aspects: 6 x 4 =24; IRT scaled to 0 to 30 point reporting scale	Max Raw score: average task score ±3 points: 35 tasks x ± 3 points +105 points; IRT scaled to a 10 to 90 point reporting score
1. Informal introductions & some everyday questions 2. Monologue based on task card 3. Discussion based on topic in task 2	1. Two everyday questions 2. Two questions based on combined information from heard and read texts on university life and academic topics. 3. Two questions based on discussion of university life and lecture excerpts	1. Six tasks: Speaking & Reading: Read aloud written language texts of 40 to 60 words on academic topic 2. Ten tasks: Listening & Speaking: Repeat sentence of typical spoken language 3. Six tasks: Speaking: Describe image, (semi)academic topics 4. Three tasks: Listening & Speaking: Re-tell lecture on academic topic 5. Ten Tasks: Listening & Speaking: Answer short question on university life and academic topics

TABLE 1
Assessment of spoken skills in IELTS, TOEFL iBT, and PTE Academic

IELTS	TOEFL iBT	PTE Academic
Two task types (both independent); one task from each; total of 2 tasks	Two task types (one independent, one integrated); one task from each; total of 2 tasks	6 task types (one independent, five integrated); total of 15 tasks
Max raw score: 2 tasks, each rated on 4 'criterion areas' using 9 band scales: 2 x 4 x 9 = 72; arithmetically reduced to a band 0 to 9 writing score: raw score divided by 8 = reported band score	Max raw score: 2 tasks each rated on 5 point rating scale: 2 x 5 = 10; IRT scaled to a 0 to 30 point reporting scale	Max Raw score: points added over all items and item types; average score per item: 6 points; 15 tasks x ±6 = ±90 points; IRT scaled to a 10 to 90 point reporting score
1. Describe information from graph, chart, diagram

2. Write essay on provided topic | 1. Answer a question based on information read and heard on academic topic

2. Answer a question based on own knowledge and experience | 1. Two tasks: Reading & Writing and Summarize written text in a single sentence

2. Two tasks: Listening & Writing and Fill in the blanks. The stimulus presents a real-life, authentic audio excerpted from an academic lecture. The stimulus also presents a reading text which is a transcription of the audio recording with 4-7 words missing from the text

3. Two tasks: Listening & Writing and Write from dictation: The stimulus presents a short sentence of 8-11 words on academic topic

4. Two tasks: Listening & Writing and Summarize spoken text on academic topic. Word limit: 50-70 words.

5. One task: Write essay. Test-takers express their views on a general academic topic. Test-takers write a persuasive essay and support their position or opinions with details and examples. Word limit: 200- 300 words

6. Five tasks: Reading & Writing and Fill in the blanks; the stimulus presents a real-life gapped reading text from an academic source. The text is 100-200 words long. It is continuous (written in paragraphs) and coherent |

TABLE 2
Assessment of written skills in IELTS, TOEFL iBT, and PTE Academic

Every measurement in an assessment contains a degree of error, referred to as the "error of measurement". The true score of a test-taker is therefore within a range of scores around the reported score. The extent of that range is determined by the size of the error of measurement. For example, if the reported score is 60 and the error of measurement is 3, then the true score, with 68% certainty, is within one measurement error from the reported score; that is within the range of 57 (60-3) and 63 (60+3). The true score, with 95% certainty, is within twice the measurement error; that is within the range of 54 (60-2x3) to 66 (60+2x3).

PTE Academic has been carefully compiled to maximize score comparability between the four skills using parallel test information functions. In Classical Test Theory, the error of measurement is an estimate of the error at any point on the scale; it is calculated without taking into account whether the score is low, medium or high. Based on data from 30,000 test-takers, the Standard Error of Measurement (SEM) on PTE Academic is 2.32. By contrast, in Item Response Theory, the error of measurement is not uniform on the scale. The Conditional Error of Measurement (CEM) on PTE Academic overall score varies from 2.5 to 3.5 depending on the CEFR level, with error at 2.41 being lowest for scores in the B1 range (GSE 40-58) and below 3 from A1 to C1. Based on the CEM, test reliability in the GSE range 53-79 (the most relevant for tertiary education admission purposes) is calculated to be 0.97 for the overall score and 0.91 for all subskills except for listening which is at 0.92. Figure 4 below compares the error of measurement expressed on the GSE for PTE Academic with those transformed to the GSE for TOEFL iBT and IELTS.

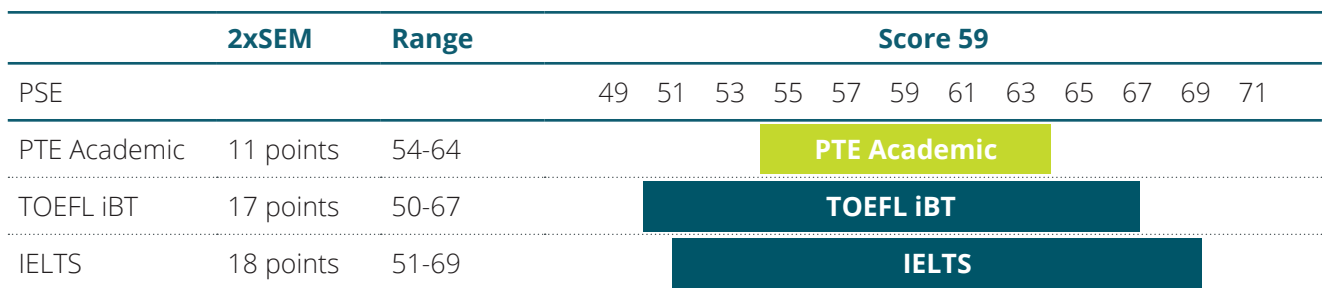|	2xSEM	Range	Score 59											
PSE			49	51	53	55	57	59	61	63	65	67	69	71
PTE Academic	11 points	54-64				PTE Academic								
TOEFL iBT	17 points	50-67			TOEFL iBT									
IELTS	18 points	51-69			IELTS									

FIGURE 4

Error of measurement for PTE Academic, TOEFL iBT, and IELTS expressed on the GSE

Conclusions

The Global Scale of English offers a granular descriptive system to define graded learning, teaching and assessment objectives. It is mathematically defined as a linear transformation of the scale underlying the levels of the CEFR.

Empirically based estimates of concordance between the three major English tests used for the admittance requirements of foreign students are available from previously reported studies.

Evidence of the reliability and validity of the GSE scale is presented and an explanation for the higher reliability of PTE Academic vs. IELTS and TOEFL is offered based on an analysis of the tests' composition. Assertions about higher reliability of PTE Academic are also supported by evidence on comparing test repeater data from PTE Academic and TOEFL (no such data are publically available for IELTS).

References

Barkaoui, K. (forthcoming). *Examining Changes over Time in Repeaters' L2 Writing Scores: The Case of PTE-Academic Writing Section*.

Bayliss, A., & Ingram, D. E. (2006). *IELTS as a Predictor of Academic Language Performance. Australian International Education Conferenc*e. Retrieved September, 12, 2016, from http://aiec.idp.com/uploads/pdf/ BaylissIngram%20(Paper)%20Wed%201630%20MR5.pdf

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: some evidence from American Universities. In *Language Testing*, 29, 3, pp. 421-442

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

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual.* Strasbourg: Language Policy Division. Retrieved September, 15, 2016, from https://www.coe.int/t/dg4/linguistic/ Source/ManualRevision-proofread-FINAL_en.pdf

De Jong, J.H.A.L. (2009a). *Unwarranted claims about CEF alignment of some international English language tests*. Paper presented at EALTA Conference, June 2009. Retrieved May, 25, 2015, from http://www.ealta. eu.org/conference/2009/docs/friday/John_deJong.pdf

De Jong, J.H.A.L. (2009b). *Dimensionality and the Construct of Language Proficiency.* Paper presented at Language Testing Forum, November 2009

De Jong, J.H.A.L., & Zheng, Y. (2016). Linking to the CEFR: validation using a priori and a posteriori evidence. In: Banerjee, J., & Tsagari, D. (Eds), *Contemporary Second Language Assessment.* London: Bloomsbury Academic, pp. 83-100

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education, In *Melbourne Papers in Language Testing*, 2, pp. 68-85

ETS (2015). *Linking TOEFL iBT ™ Scores to IELTS® Scores – A Research Report.* Retrieved September, 25, 2016, from http://www.ets.org/s/toefl/pdf/ linking_toefl_ibt_scores_to_ielts_scores.pdf

Graham, J.G. (1987). English Language Proficiency and the Prediction of Academic Success. In *TESOL QUARTERLY*, 21, 3, pp. 505–521

IELTS (n.d.). *Common European Framework. How should the CEFR be used by recognising institutions wishing to set language ability requirements?* Retrieved November, 2, 2016, from https://www.ielts.org/ielts-for- organisations/common-european-framework

Liao, C. W, Qu, Y, & Morgan, R. (2010). *The Relationships of Test Scores Measured by the TOEIC Listening and Reading Test and TOEIC Speaking and Writing Tests.* TOEIC Compendium Report TC-10-13. Princeton, NJ: ETS. Retrieved September, 25, 2016, from https://www.ets.org/toeic/research

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency.* Berlin, New York: Peter Lang

Pearson (2010). *Aligning PTE Academic test scores to the Common European Framework of Reference for Languages.* Retrieved October, 28, 2016, from http://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_ Scores_CEF.pdf

Reckase, M., & Xu, J. (2015). The evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. In *Educational and Psychological Measurement,* 75, 5, pp. 805-825

Spolsky, B. (1997). The ethics of gatekeeping tests: what have we learned in a hundred years? In *Language Testing*, 14, 3, pp. 242-247

Tannenbaum, R., J., & Wylie, E. C. (2008). *Linking English-Language Test Scores onto the Common European Framework of Reference: An Application of Standard-Setting Methodology*. Princeton, NJ: ETS. Retrieved September, 25, 2016, from http://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Taylor, L. (2004). *IELTS, Cambridge ESOL examinations and the Common European Framework.* Research Notes, Issue 18, November, pp. 2-3. Retrieved December, 13, 2016, from http://prd.cambridgeesol.org/images/23135-research-notes-18.pdf

Woodrow, L. (2006). Academic Success of International Postgraduate Education Students and the Role of English Proficiency. In *University of Sydney Papers in TESOL,* 1, pp. 51-70

Zhang, Y. (2008). *Repeater Analyses for TOEFL® iBT RM-08-05*. Retrieved October, 28, 2016 from https://www.ets.org/Media/Research/pdf/RM-08-05.pdf

Zheng, Y., & De Jong, J.H.A.L. (2011). *Establishing construct and concurrent validity of Pearson Test of English Academic* (1-47). Retrieved October, 25, 2016, from http://pearsonpte.com/wp-content/uploads/2014/07/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf

Glossary

Term	Definition
CEFR	Common European Framework of Reference for Languages
concordance table	A table showing how scores on one test relate to scores on another test.
correlation	A statistic showing the interdependence between two variables.
field testing	A method used to gather data from a group of people who represent the target test-taking population in order to calibrate a test.
GSE Learning Objective	A description of what a student can do at a particular point on the Global Scale of English.
IRT	Item Response Theory: A method used to apply a mathematical model to test data. It predicts the probability of an item being correctly answered based on the mathematical function of the ability of the person and the difficulty of the item.
logit scale	A standardised scale of logarithmic units.
regression function	A mathematical function expressing the relation between a dependant variable (y-axis) and an independent variable (x-axis).
reliability	The consistency of the measures of a test.
standard deviation (SD)	A statistic showing the amount of variation in a data-set. An SD close to 0 means the data point is close to the mean.
transformation function	A mathematical function enabling the transformation of values on one scale to corresponding values on another scale.
unidimensionality	A property of a test characterized by item homogeneity.

Appendix

Alignment of scores of some major English Proficiency Tests to the CEFR

CEFR	GSE	PTE Academic	TOEFL iBT (1)	TOEFL PBT	IELTS (2)
C2	85-90	85-90	no score	no score	8.5-9.0
C1	76-84	76-84	110-120	635-677	7.5-8.5
B2	59-75	59-75	87-109	569-634	6.5-7.5
B1	43-58	43-58	57-86	490-568	5.5-6.5
A2	30-42	30-42	no info	no info	4.5-5.5
A1	22-29	22-29	no info	no info	no info
<A1 Tourist	13-21	13-21	no info	no info	no info
<Tourist	10-12	10-12	no info	no info	no info

No score: Indicates that for this test there are no scores above C1

No info: Indicates that for this test we had no data at this level to base the concordance on

(1) According to the original, methodologically correct version published in: Richard J. Tannenbaum and E. Caroline Wylie (2008). Linking English-Language Test Scores onto the Common European Framework of Reference: An Application of Standard-Setting Methodology. ETS, Princeton, NJ. Retrieved December, 13, 2016 from https://www.ets.org/Media/Research/pdf/RR-08-34.pdf. In 2015 a new claim about alignment was made in Spiros Papageorgiou, Richard J. Tannenbaum, Brent Bridgeman and Yeonsuk Cho (2015). The Association between TOEFL iBT® Test Scores and the Common European Framework of Reference (CEFR) Levels. Research Memorandum RM–15-06. ETS, Princeton, NJ. Retrieved December, 13, 2016 from https://www.ets.org/Media/Research/pdf/RM-15-06.pdf. This claim sets the CEFR cut-off at twice the error of measurement below the observed cut-off (ETS RM-15-06, page 8), which effectively means that 97.5% of the learners assigned to a level are in fact more likely to be one level below. We therefore do not present this view in the table.

(2) The IELTS – CEFR relationship provided by the makers of IELTS (http://www.ielts.org/researchers/common_european_ framework.aspx) is understood to be overrated. Therefore the relation in this table is based on research studies by ETS and Pearson: ETS (2010). Linking TOEFL iBT Scores to IELTS Scores – A Research Report. Retrieved December, 13, 2016 from https:// www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores. pdf; Pearson (2012) PTE Academic Score Guide. Retrieved December, 13, 2016 from http://pearsonpte.com/wp-content/ uploads/2014/07/PTEA_Score_Guide.pdf.