PEARSON ALWAYS LEARNING

Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages

Introduction

Pearson Test of English Academic (PTE Academic) is a new computer-based international English language test. Pearson developed PTE Academic in response to demand from higher education, government and other customers for a test that will more accurately measure the English communication skills of international students in an academic environment. The purpose of this test is to measure test takers' academic English language competency in Listening, Reading, Speaking and Writing. PTE Academic is endorsed by the Graduate Management Admission Council® (GMAC®). GMAC is the owner of the Graduate Management Admission Test ® (GMAT®).

To develop this comprehensive new computer-based English language test, Pearson worked with internal and external test development experts (see acknowledgement). In addition, the company conducted an extensive field test programme to test the items of PTE Academic and measure their effectiveness in assessing a test taker's ability to communicate in English in an academic environment. This document provides an overview of the process that was undertaken to provide evidence for the relation of PTE Academic scores to the levels of the Common European Framework of Reference for Languages (Council of Europe, 2001).

PTE Academic is a multi-level, integrated-skills test of English language proficiency. It is designed to assess English language competence set in the context of academic programmes of study that are available around the world. The development of the tasks and items for the test followed consultation with external stakeholders and test development professionals. PTE Academic is supported by two external advisory boards composed of experts in applied linguistics and assessment who have overseen the development of the test from a professional perspective.

PTE Academic uses 20 item types reflecting different modes of language use and setting different response tasks or response formats. The maximum duration of the test is three hours. The test is administered entirely on computer in secure test centres using Pearson's state-of-the-art security measures (Lopes, 2010).

PTE Academic scores are delivered online, typically within five business days (current average is 2 days). The score report is made available to test takers via their personal login and to registered institutions via their secure login. The score report provides three types of scores: an Overall Score, scores for Communicative Skills (i.e. Listening, Reading, Speaking and Writing) and scores for Enabling Skills (i.e. Grammar, Oral Fluency, Pronunciation, Spelling, Vocabulary and Written Discourse). The score scale ranges from 10 to 90.

The Common European Framework

For reasons of transparency it is useful to relate numerical test scores to a descriptive system which facilitates interpreting test scores in terms of predicted potential for behaviour of test takers. In the context of language testing stakeholders can be considered fortunate in that they have access to a descriptive system that has wide recognition. The Council of Europe published the Common European Framework of Reference for Languages in 2001. The Framework (abbreviated as CEF or CEFR, never CEFRL) was the product of almost twenty years of cooperative work by language teachers and experts from all member states of the Council of Europe with representation from Canada and the USA. It was initiated by the Committee of Ministers from the member states who realized the importance of language learning, teaching and assessment in a world that was rapidly becoming a global society (Council of Europe, 1982). In order to stimulate language learning and enhance the usefulness of the effort of learning they considered it important to establish a means to enable exchange of information about the goals and the products of learning and teaching languages and to define these in terms of potential for using language for communication.

Part of the descriptive framework is the definition of a series of ascending levels for describing learner progress. The definition is based on more than 500 descriptors of language activities defining both the content and the quality of those activities. The descriptors have been calibrated in a project funded by the Swiss Government (North, 2000). After calibration the scale underlying the descriptors was divided into six levels that were considered meaningful in the context of communicative language use. These levels are indentified by a letter and number, where the letters A, B, C refer to Basic, Independent and Proficient language use respectively and each of these three levels is further divided in a lower and higher level by adding either 1 or 2 to the letter, resulting in a six level system going from A1 to C2. The system allows for refinement by further dividing each of the levels, e.g., A1 can be subdivided in A1.1 and A1.2.

Within this system the level B1 was originally labelled the “Threshold” (Van Ek, 1974: Van Ek and Trim, 1990) level expressing its function as a minimum competency level to be able to use a language as independent agents (without help) in dealing with other speakers of that language. Both A levels (A1 and A2) define levels that allow for functional communication provided support is available to the language user, for example by using simplified language, speaking slowly, etc.

From its definition (see Exhibit 1) it can be concluded that the level B2 is required to be likely to function successfully in language exchange as one may encounter in higher education. Basically this implies that students who have attained level B2 in a foreign language which is used as the language of instruction and communication in an institution for tertiary education would not be disadvantaged significantly because of the language in comparison to students for whom that language is their first language.

> Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and Independent disadvantages of various options.

Exhibit 1: Global descriptor for Level B2 © Council of Europe 2001

In order to assist users of the CEF in relating tests or exams to the descriptive system of levels, the Council of Europe has published a manual (Council of Europe, 2009). This manual distinguishes four necessary stages in building an argument for relating test scores to the descriptive levels of the CEF:

1. Familiarization: persons involved in the process of relating the test to the CEF should have a thorough knowledge of the CEF;

2. Specification: ascertaining whether the test provides sufficient coverage of the framework as described mainly in chapters 4 and 5 of the CEF publication;

3. Standardization: ascertaining whether judges or raters involved in evaluating the difficulty of test tasks and/or the proficiency of test takers performing these tasks are well equipped to relate perceived difficulty of tasks and proficiency of test takers to the descriptive levels of the CEF;

4. Empirical validation: ascertaining (a) whether the test itself meets requirements of reliability and validity to serve as a measurement instrument and (b) whether the relation with the CEF can be supported by statistical data.

Familiarization, Specification and Standardization

PTE Academic differs from most tests and exams for which a relation with the CEF is claimed in that the test was designed to measure language competence according to the principles of the CEF and to address specifically language competencies in the range from upper B1 to lower C2. During the training of item writers, item reviewers and human raters the stages of Familiarization, Specification and Standardization were each addressed consecutively, culminating in actual exams assessing the consistency and agreement on level specification and evaluation. The three separately defined stages in the linking process need each to be addressed when an existing test is to be linked to the CEF post-hoc. For PTE Academic they were addressed as part of the development process.

Prior to each stage of the development of PTE Academic all individuals involved at that stage received intensive training in understanding and using the CEF. In particular the stages of test specification and the selection and definition of language tasks to be included in PTE Academic were based on the CEF. A checklist was used to assess adherence to the considerations presented throughout the CEF publication. This “Checklist of Considerations” was drafted by gathering all 144 considerations from the CEF. To the right of each consideration two checkboxes were provided, one to assess whether the specific consideration was *applicable* in the context of the test and its intended use, and a second box to be checked when the consideration was actually *applied*. Exhibit 2 shows the first three of these considerations with the checkboxes as they were used during the test development process. Whenever considerations were considered applicable it was made sure they were also applied.

Nr. consideration	Applicable	Applied
Users of the framework may wish to consider and where appropriate state:		
1. to what extent their interest in levels relates to learning objectives, syllabus content, teacher guidelines and continuous assessment tasks (constructor-oriented).		
2. to what extent their interest in levels relates to increasing consistency of assessment by providing defined criteria for degrees of skill (assessor-oriented).		
3. to what extent their interest in levels relates to reporting results to employers, other educational sectors, parents and learners themselves (user-oriented).		

Exhibit 2: Sample from 'Checklist of Considerations'

Once the test and task specifications were finalized and approved by the external Technical Advisory Group, training sessions were organized for local groups of item writers in Australia (Sydney), Europe (UK, London) and the USA (Washington). The training included an in depth introduction to the CEF followed by special attention to the assessment aspects dealt with in the CEF and was rounded off by training addressing how to understand the principles of the level descriptors. Item writers were familiarized by using self-assessment grids from the CEF to assess their own level of competence in two or three foreign languages. Next they were required to challenge each others' claims and defend their own assessment. This exercise was followed by activities in pairs ordering blind versions of CEF descriptors. Next sample items and sample responses were discussed and assigned to levels on the CEF scale. The training was concluded with individual activities applying the CEF scale to item selection and test taker response assessments.

Item writers were tasked to address the language tasks relevant in university or higher education settings, covering all four skills and addressing CEF levels from upper B1 to lower C2. They were required to use real texts and realistic tasks. They had to name their source so Pearson could ask the copyright holders for permission to use their material. For each item the item writers also had to indicate what CEF level they were aiming their item to address.

In the rounds of item reviewing and revision item reviewers were again asked what CEF level they thought the item was assessing. Item writers' and item reviewers' judgements were recorded and stored with the items in the item bank.

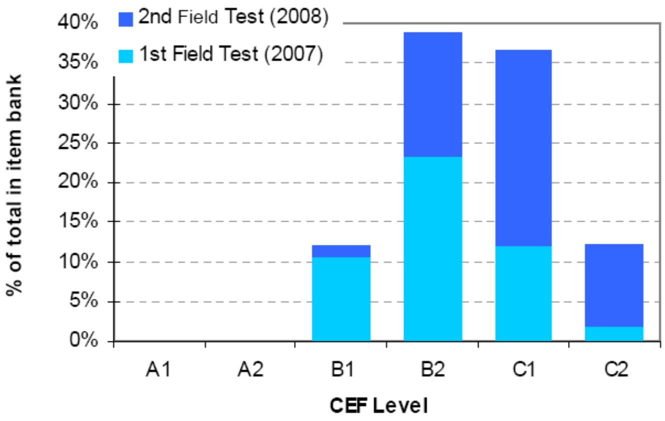

Figure 1: Distribution of items over targeted CEF levels

As part of the development process two field test events were organized. The second, apart from increasing the volume of items in the item bank, also served to specifically address language tasks and CEF levels that were found to be insufficiently represented in the first round of field testing. For example, Figure 1 shows how the distribution of items over the CEF levels was corrected in the second field test.

Empirical Validation

Validity and reliability of PTE Academic scores

A condition to be able to link the PTE Academic scores to the levels of the CEF is that the scores provide a reliable and valid assessment of test takers' abilities. In general terms:

Validity is the extent to which the test results are relevant and meaningful for the intended use of the test.
Reliability is the extent to which the test results can be relied upon, i.e., that the results will be similar on repeated occasions.

There are many aspects to validity and fully addressing all these aspects would exceed the limits of this document. One major aspect of the validity of a test purporting to assess the competencies of test takers to use a language which is not their first language is its power to distinguish them from peers for whom the language tested is their first language. To assess this quality the test taker samples to whom every field tested item was administered included between 10 and 15 per cent of native speakers of comparable age and educational background as the target population of foreign language test takers. Figure 2 presents the self-reported age of test takers plotted against their z-score[1] on the total field test version of PTE Academic. Test takers for whom English is their first language have been plotted with an open red circle whereas test takers for whom English is a foreign language a solid blue dot is used.

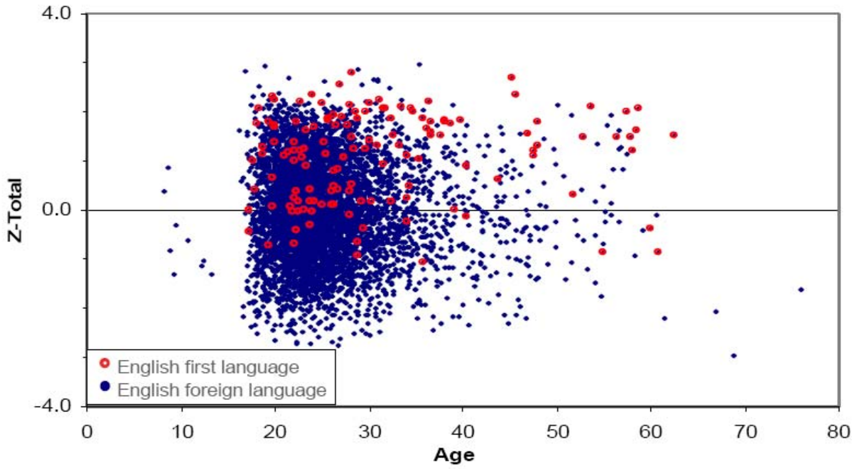

Figure 2: Age vs. z-scores for test takers with English as a first and foreign language

Figure 2 shows that test scores are generally above the mean for first language test takers whereas foreign language test takers are evenly distributed over the z-scale. In addition Figure 2 shows that this differential scoring pattern occurs irrespective of age.

[1] Z-score expresses scores as distance from the mean score, where the mean score is set to 0 and the standard deviation is set to 1. Negative z-scores therefore represent scores below the mean and positive z-scores represent scores above the mean.

Other evidence of validity was gathered by independent external researchers. Dr. Kieran O'Loughlin from Melbourne University studied the lexical validity and found evidence for the academic nature word use in the prompts and in the responses from test takers. He also found a relation between the frequency of academic level word usage and the CEF level assigned to test takers by human raters.

Professor Fred Davidson from the University of Illinois at Urbana-Champaign conducted a sensitivity review. A small proportion of items were identified as potentially sensitive for particular groups of test takers. For a limited number of those items statistical evidence of such bias was indeed found. Items found to be sensitive were removed from the item bank.

Evidence of the reliability of PTE Academic test scores was found in the field test data by separately calibrating all odd and all even items. High correlations were found between the scores based on these test halves indicating that random draws of items from the PTE Academic item bank yield similar results. This reliability index can be seen as a post-hoc estimate of reliability. For actual live testing an a-priori approach to reliability is used. Test forms are drawn from the item bank in a stratified random sampling procedure. Selections are stratified according to (a) aspect to be measured, (b) item difficulty, and (c) length of time. The selection according to difficulty ensures the selection meets target test information functions for each of the four skills (listening, reading, speaking and writing). Meeting these target information functions results in pre-determined maxima for the measurement error along the score reporting scale and therefore an a-priori defined estimate of test score reliability. Table 1 provides the reliability estimates of the Overall Score and the Communicative Skills scores within the PTE Academic score range of 53 to 79, which is the most relevant area for admission decisions.

Score	Reliability
Overall	0.97
Listening	0.92
Reading	0.91
Speaking	0.91
Writing	0.91

Table 1: Reliability estimates for scores in the range 53–79

Statistical linking procedures

Statistical procedures for relating PTE Academic scores to the levels of the CEF scales involved both a test taker-centred and item-centred approach. For the test taker-centred approach Pearson used test taker responses three item types: Writing essay, Oral description of an image and Oral summary of a lecture. These responses were rated on the CEF scale by two human raters, independently of the ratings produced obtaining item scores. Given the probabilistic and continuous nature of the CEF scale adjacent scores were considered acceptable. The diagram in Figure 3 shows that ratings assigned to response **Y** and **Z**, though in two adjacent CEF levels (B1 and B2) in fact represent a higher level of agreement than the two rating assigned to responses **X** and **Y**, though these have been assigned the same CEF Level (B1). When CEF ratings were more than 1 level apart a third rating was called. The two closest were kept.

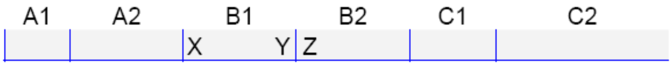

Figure 3: Agreement and adjacency of ratings on the CEF scale

The relation between ability estimates based on scored responses on PTE Academic and the CEF was represented in box-plots (Figure 4).

These box plots show substantial overlap across adjacent CEF categories, as well as an apparent ceiling effect at C2 for writing. CEF levels however are not to be interpreted as mutually exclusive categories. Language development is continuous, not in stages. Therefore the CEF scale and its levels should be interpreted as probabilistic: learners of a language are estimated to most likely be at a particular level, but this doesn't reduce to zero their probability to be at an adjacent level.

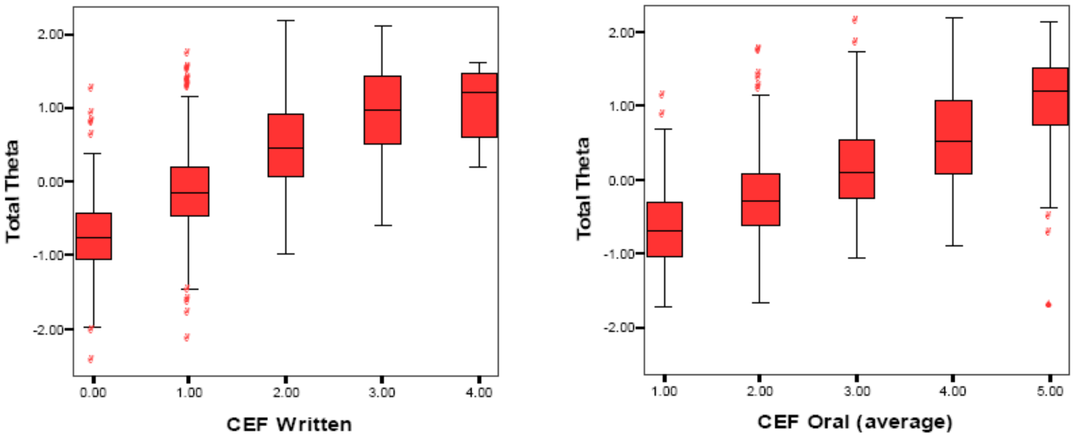

Figure 4: CEF level distribution box plots

Though the official CEF literature does not provide information on required minimum probability to "be of a level", the original scaling of the levels (North, 2000) was based on the Rasch model and cut-offs were defined at 0.5 probability. The distance between levels implies that typically anyone reaching a probability of around 0.8 to be at level x, has .5 probability of being at level x+1 and is therefore exiting level x and entering levex+1. Having a probability of 0.5 of being at level 1 implies a probability of 0.15 to be at level x+1 and a little as 0.05 at level x+2.

Therefore the overlap shown in Figure 4 corresponds to the modelled expectation.

In order to estimate the cut-offs for the CEF ratings Pearson used the computer program FACETS (Linacre, 1988; 2005) defining four facets as shown in Table 2.

Facet Nr	Facet	n
1	Candidates	4028
2	items	94
3	Skills (Oral and Written)	2
4	Raters	147

Table 2: Data definition for CEF FACETS analysis

Figure 5 shows the probability curves for the rated categories ‘below A2’[2], ‘A2’, ‘B1’, ‘B2’, ‘C1’ and ‘C2’.

All category curves are most probable for some range of the CEF theta; ranges tend to become narrower towards the end of the scale. This is in agreement with the original CEF-scaling proposed by North (2000), although category 3 (=B2) is slightly narrower than expected.

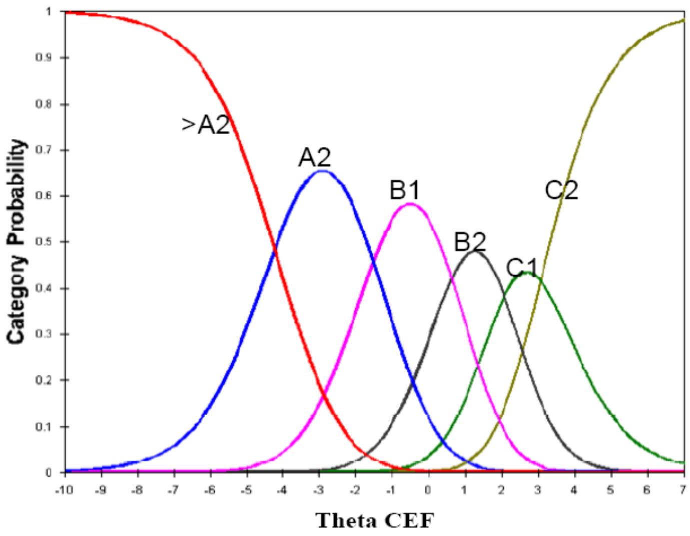

Figure 5: Probability Category Curves

Figure 6 shows modelled category expectation (red curve) for ranges of CEF theta category as well as the observed average data for groups of size + 100 (black crosses) In addition the error on either side of the modelled expectation is indicated by thin grey lines. It is clear that the data generally fit the model though there is some noise at the very low end. As expected the error becomes larger at the upper extreme.

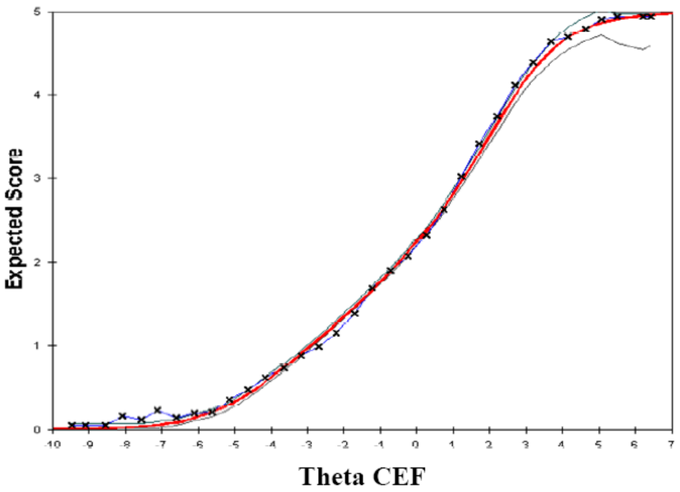

Figure 6: Modelled and observed categorization in relation to Theta CEF

[2] No distinction is made at proficiency levels below A2 because the test does not aim to measure at these initial learning levels.

The estimates of category boundaries on the CEF theta scale are shown in Table 3.

Category	CEF Level	Theta CEF (Lower bounds)
0	BELOW A2	N/A
1	A2	-4.24
2	B1	-1.53
3	B2	0.63
4	C1	2.07
5	C2	3.07

Table 3 Category lower bounds on theta CEF

The relationship between the scaled CEF ratings and the Theta PTE for all candidates with information on both scales (n=3318) is shown in Figure 7. The correlation between the two measures is 0.69. A better fitting regression is obtained with a first order polynomial (curved red line), yielding an r2 of slightly over 0.5.

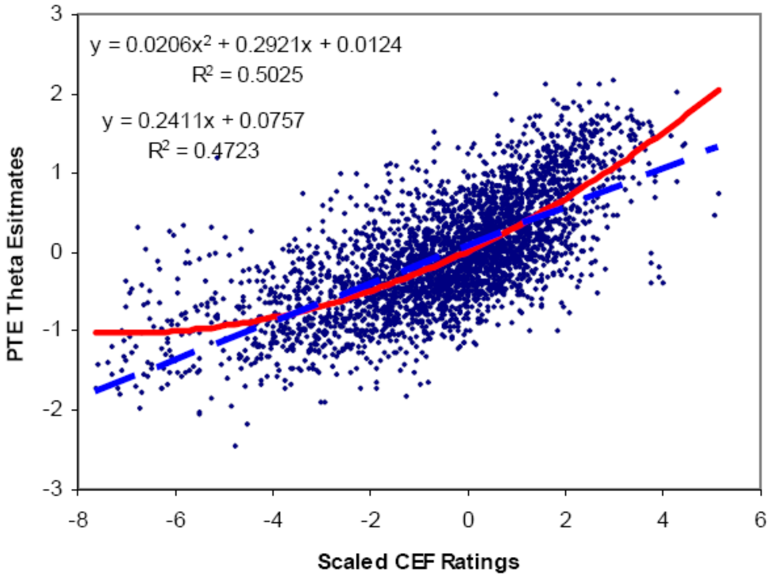

Figure 7: Relation between Theta CEF and Theta PTE

Because of noisy data at the bottom end of the scales, the lowest performing 50 candidates were removed. Further analyzes were conducted with the remaining 3268 subjects. Figure 8 shows the cumulative frequencies for these 3268 candidates for whom theta estimates are available on both scales. The cumulative frequencies are closely aligned, although the PTE scale clearly shows smaller variance.

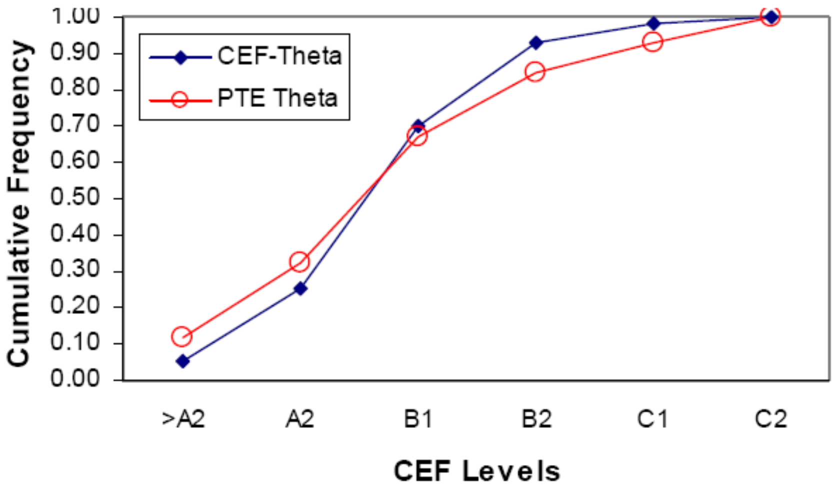

Figure 8: Cumulative Frequencies for CEF Levels on CEF and PTE theta scales (n-3318)

In the next stage an equipercentile equation was chosen to express the CEF lower bounds on the PTE theta scale. The cumulative frequencies are shown in Figure 9, and the projection of the CEF lower bounds on the PTE theta scale together with the observed distribution of field test candidates over the CEF levels is shown in Table 4.

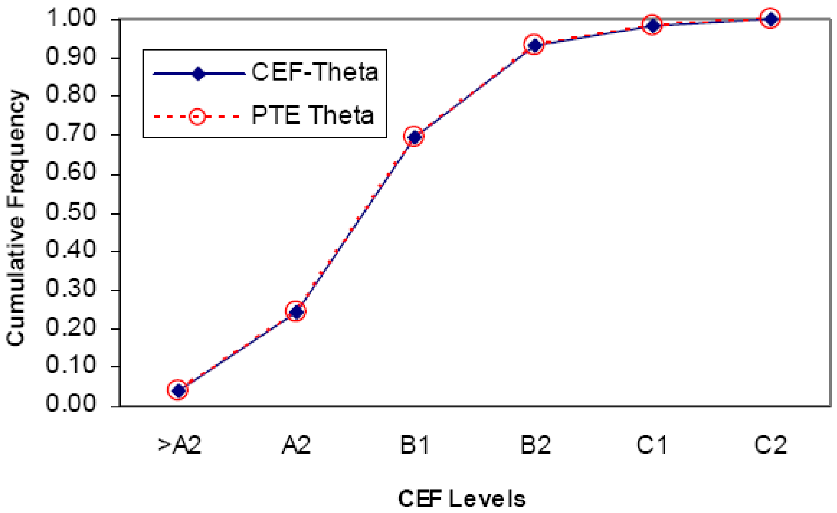

Figure 9: Cumulative frequencies on CEF and PTE theta scales after equipercentile equating

CEF Levels	Theta PTE	Frequency	Percentage	CumFreq
>A2	-1.366	126	4%	0.04
A2	-1.155	677	21%	0.25
B1	-0.496	1471	45%	0.70
B2	0.274	769	24%	0.93
C1	1.105	170	5%	0.98
C2	>1.554	55	2%	1.00
Totals		3268	100%	

Table 4: Final Estimates for CEF lower bounds on PTE theta scale

As expected, given the descriptor for CEF level C2, (see Exhibit 3) the number of candidates achieving this level is very low.

> Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Exhibit 3: Descriptor for CEF Level C2 © Council of Europe 2001

At item development stage item writers indicated for each item which level of ability expressed in terms of the CEF levels they intended to measure, i.e., did they think test takers would need to be able to correctly solve the items. Table 5 provides the mean observed difficulty for each of the CEF levels targeted by the item writers.

Intended CEF Level	Mean observed difficulty
A2	0.172
B1	0.368
B2	0.823
C1	1.039
C2	1.323

Table 5: Intended and observed item difficulty

Figure 10 shows the estimated lower bounds of the difficulty of items targeted at each of the CEF levels plotted against the lower bounds of these levels as estimated from the independent CEF ratings of test takers' responses by human raters. Both estimates, derived independently, agree to a high degree. (r=0.99).

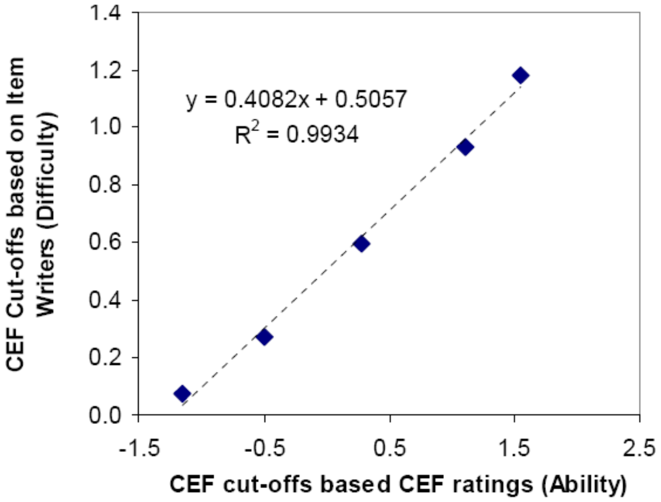

Figure 10: Lower bounds of CEF levels based on targeted item difficulty versus lower bounds based on CEF ratings of candidates' responses

What It Means to be ‘At a Level’

Score users should be aware that test providers may use different methods to relate scores on their tests to the CEF scale. If test providers make no or insufficient information publically available, the validity of alignment claims cannot be verified and their meaning cannot be interpreted. Alignment claims are irrelevant if the test provider does not explain what it means to be ‘at a level’ (See Adams and Wu, 2002, pp 197-199).

The premise underlying Pearson’s alignment study is that the ability required to stand a reasonable chance at successfully performing *any* of the tasks defined at a particular CEF level is the ability needed to successfully perform the average task at that level. For B1, for example, this average is the average over *all* B1 tasks, ranging from the easiest B1 task to the most difficult. As learners develop their ability at B1 level the probability of them successfully performing any B1 task grows and ultimately reaches a point where there is a reasonable chance at performing the average task at B2 successfully, i.e., they have entered the B2 level.

By contrast, some users of the CEF scale base their score equivalence on the premise that learners who stand a reasonable chance at performing the *easiest* task at a particular CEF level are able to function at that level. Within such an interpretation reaching any of the CEF levels is obviously less demanding than in Pearson’s definition.

On the other hand, the document *Self-assessment Checklist*[3], on the Council of Europe’s website, provides lists of descriptors (can-do statements) for each of the levels. Users are told to tick the statements they feel they can do under normal circumstances. On the A1 page the document then states “If you have over 80% of the points ticked, you have probably reached Level A1.” This interpretation of what it takes to be at a level is obviously too demanding as it would leave very little room for learners to grow within a level.

To sum up, claims about alignment with the CEF can only be properly interpreted if they are accompanied by information revealing the underlying premise about what it means to be “at a level”: likely to be successful with tasks at the bottom of a level, standing a fair chance to succeed on any task, or able to perform almost all tasks? The table below shows for each of the CEF levels A2 to C2 which PTE Academic scores predict likelihood of successful performance on the easiest, the average and the most difficult tasks within each of the CEF levels.

CEF level	Easiest	Average	Most difficult
C2	80	85	n.a.
C1	67	76	84
B2	51	59	75
B1	36	43	58
A2	24	30	42

Table 6: PTE Academic scores predicting likelihood of successful performance on CEF level tasks

Pearson’s definition of “being at a level” is shown in the middle column, i.e., learners who are likely to be successful on any task at that level.

From discussion with the UKBA, it has become apparent that the UKBA requires students to demonstrate ability to perform tasks at the lower bounds of the B1 and B2 levels as defined by the Common European Framework (CEF) rather than the ability to cope with the complete range of tasks at those levels. Consequently, to meet the UKBA requirements of being at a particular level the PTE Academic scores presented in the left hand column need to be attained.

[3] http://www.coe.int/T/DG4/Portfolio/documents/appendix2.pdf; retrieved 11/05/2011

Figure 11 shows the relationship between a test taker's ability in terms of likely performance on B1 tasks and the variation in difficulty of B1 tasks projected onto the PTE Academic score scale. A score of 36 marks the lower boundary and means that a candidate will be able to perform the easiest tasks at B1. A score of 43 means a test taker will be able to perform the average tasks at B1, and 58 represents mastery of Level B1.

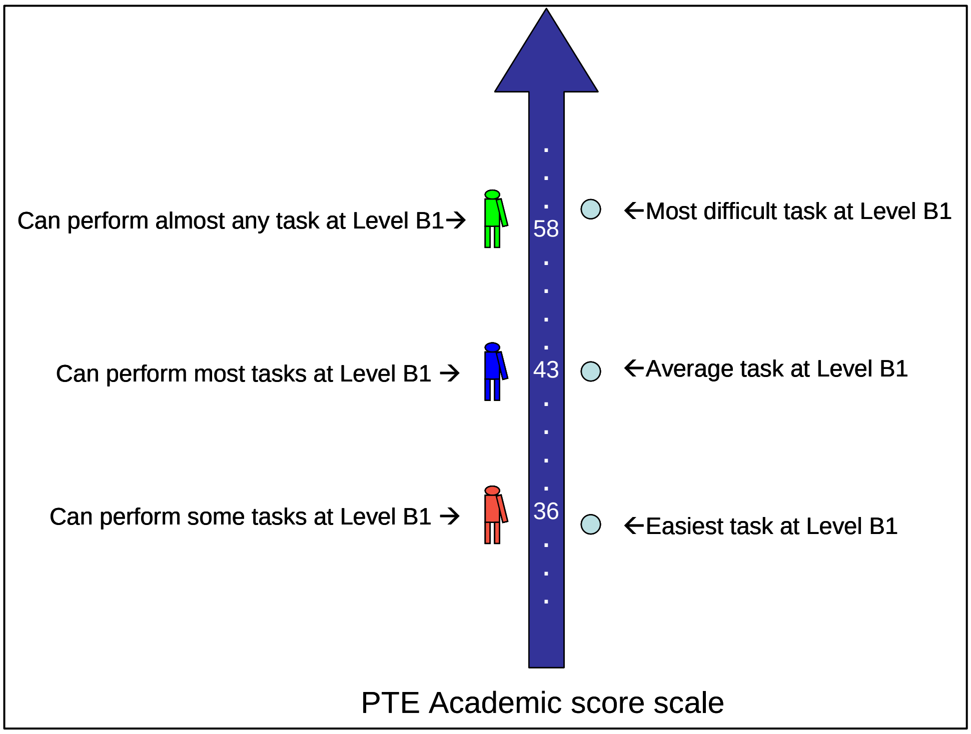
Can perform almost any task at Level B1→
Can perform most tasks at Level B1 →
Can perform some tasks at Level B1 →
58
43
36
←Most difficult task at Level B1
←Average task at Level B1
←Easiest task at Level B1
PTE Academic score scale

Figure 11: Relation between test taker ability, task difficulty and the PTE Academic score scale at the B1 Level

Score Correspondence with Easiest CEF Level Tasks

The CEF levels are based on the calibration of descriptors of task performance using the one-parameter Rasch model. The Rasch model is an Item Response Theory (IRT) model which models the probability of a response being correct or incorrect given the ability (β) of a person and the difficulty (δ) of an item.

In the case of person *n* responding to item *i*, the probability that the response will be correct (X_{ni} = 1) is given by:

$$\Pi\{X_{ni}=1\}=\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}} ,$$

where β_n is the ability of person *n* and δ_i is the difficulty of item *i*.

Pearson's definition of the minimal ability of 'at a level' is *the ability threshold at which it is more likely than not for a person to be successful in performing any task at that level*. As the number of possible tasks at any level is infinite, the best estimate of the average task of the universe of the tasks at a level is the midpoint between the lowest and the highest point of difficulty, i.e., between the difficulty at the lower thresholds of a level and the next level up.

If the minimal ability to be 'at a level' is defined as the ability threshold at which it is more likely than not for a person to be successful in performing the easiest task at that level, the minimum ability to be 'at a level' actually equals the difficulty boundary at the lower threshold of the level. Furthermore, given Pearson's placing of the lower boundaries of the CEF levels at midway between the easiest and most difficult tasks, computing the threshold at the easiest task at a level simply requires positioning at the level of difficulty of the easiest task. This by definition is half a level below the Pearson threshold.

Figure 12 illustrates the projection of cut-offs on the ability scale (A) based on the easiest Level X task as the minimal requirement and (B) as the average Level X task as the minimal requirement.

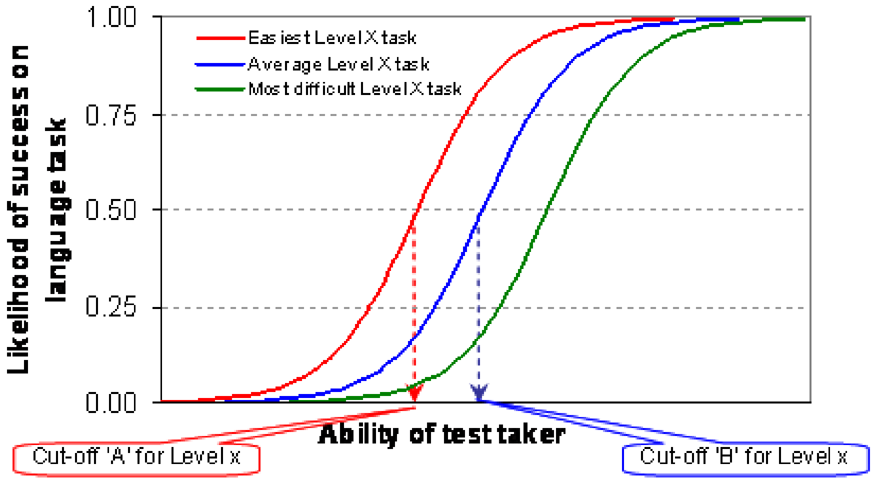

Figure 12: Projection of cut-offs on ability scale from easiest Level X task (A) and from average Level X task (B)

Interpreting PTE Academic Tier 4 Scores

The ability of test takers at the lower boundary of a level differs from the ability of test takers likely to be successful in performing the average task at a level. Descriptions of the ability at these lower boundaries are provided in Table 7 for B1 and B2.

PTE Academic Score	Common European Framework Level	What this means for a score user
51	**Lower boundary B2**	Has sufficient command of the language to deal with most familiar situations, but will often require repetition and will make many mistakes. Can deal with standard spoken language, but will have problems in noisy circumstances. Can exchange factual information on familiar routine and non-routine matters within his field with some confidence. Can pass on a detailed piece of information reliably. Can understand the information content of the majority of recorded or broadcast material on topics of personal interest delivered in clear standard speech.
36	**Lower boundary B1**	Has limited command of language, but it is sufficient in most familiar situations provided language is simple and clear. May be able to deal with less routine situations on public transport e.g., asking another passenger where to get off for an unfamiliar destination. Can retell short written passages in a simple fashion using the original text wording and ordering. Can use simple techniques to start, maintain or end a short conversation. Can tell a story or describe something in a simple list of points.

Table 7: Description of ability at the lower boundaries of CEF Levels B1 and B2 (adapted from Council of Europe, 2001)

Acknowledgements

The research reported in this paper has been conducted by Pearson Test Development and Research staff under the supervision of Prof. dr. John H.A.L. de Jong, supported by dr. Kirk Becker, Senior Research Scientist at Pearson VUE and by external researchers Prof. dr. Daniel Bolt from the University of Wisconsin at Madison and dr. Joshua Goodman, Assistant Professor at James Madison University.

The results of the linking study have been reviewed and approved by the Technical Advisory Group for PTE Academic, an independent committee consisting of Ernesto Anastasio from the Graduate Management Admission Council, Chris Caridia, freelance consultant, research and language teaching materials, Prof. dr. Alan Davies emeritus professor of the University of Edinburgh, Prof. dr. Mark Reckase from Michigan State University, and Prof. dr. Shin-ichi Mayekawa from the Tokyo Institute of Technology.

References

Adams, Ray and Margaret Wu (Eds) (2002) *PISA 2000 Technical Manual*. Paris: OECD.
Council of Europe (1982) *'Recommendation no R (82)18 of the Committee of Ministers to member States concerning modern languages'*. (Republished as Appendix A to Girard & Trim, 1988).
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP.
Council of Europe (2009) *Manual for relating Language Examinations to the Common European Framework of Reference for Languages.* Author: http://www.coe.int/t/dg4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf
Girard, D. and Trim, J.L.M. (eds.) (1998) *Project no.12 'Learning and teaching modern languages for communication': Final Report of the Project Group (activities 1982–87)*. Strasbourg, Council of Europe.
Linacre, J.M (1988; 2005) A Computer Program for the Analysis of Multi-Faceted Data. Chicago, IL: Mesa Press.
Lopes, S. (2010) Test security: defeating the cheats. *Biometric Technology Today, Vol. 18, 4*, 9-11. http://www.sciencedirect.com/science/journal/09694765
North, B. (2000) *The development of a common framework scale of language proficiency*. New York: Peter Lang.
Van Ek, J.A. (1977) *The Threshold Level for modern language learning in schools*. London: Longman.
Van Ek, J.A.; Trim, J.L.M. (1991) *Threshold Level 1990*. Cambridge: CUP.