4

Linking to the CEFR:

Validation Using *a Priori* and *a Posteriori* Evidence

Ying Zheng and John H.A.L. de Jong

ABSTRACT

Linking tests to international standards, such as the Common European Framework of Reference: learning, teaching, assessment (CEFR, Council of Europe, 2001), is a way of establishing criterion-referenced validity. This chapter reports on how CEFR scales were operationalized in practice in the course of developing the Pearson Test of English Academic. Measures to link the test to the CEFR were studied at different stages of test development. *A posteriori* statistical evidence was also collected from both field tests and live tests. Field test data were used to establish the extent to which scores from this test can be linked to the CEFR, which involved both a test-taker-centered approach and an item-centered approach.

Research Background

Achieving test validity is an essential concern in test development, particularly when a test is used for high-stakes purposes. However, as Messick (1992) commented, "Many test makers acknowledge a responsibility for providing general validity evidence of the instrumental value of the test, but very few actually do it" (p. 18). More recently, Weir (2005) reported that, while most examinations claim different aspects of validity, they often lack validation studies of actual tests that demonstrate evidence to support inferences from test scores.

Messick's (1995) unified view of validity predicated that validity is a multifaceted concept, which can only be established by integrating considerations of content, criteria, and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. It is widely recognized that the validation process should start from the very beginning of test development. Schilling (2004) maintained that, in addition to *a posteriori* validity evidence (which traditionally focused on scoring validity, criterion-related validity, and consequential validity); *a priori* validity evidence (such as test design decisions and the evidence that supports these decisions) also makes a significant contribution to the establishment of validity. Similarly, Weir (2005) highlighted the importance of *a priori* validity evidence when he stated that "the more fully we are able to describe the construct we are attempting to measure at the *a priori* stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test" (p. 18). The reason is that the statistical analysis at the *a posteriori* stage does not generate conceptual labels by themselves, and therefore, to make the scores meaningful, the test developers can never escape from the need to define what is being measured at the beginning of test development.

The Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) has had a major impact on language education worldwide. Byram et al. (2012) noted that the CEFR has not only helped develop both strategic language policy documents and practical teaching materials, but it has also become the most reliable reference for curriculum planning. Linking tests to international standards such as the CEFR is a way of establishing criterion-referenced validity. As is widely acknowledged, validation is a continuous process of quality monitoring (AERA, APA, and NCME, 2014). The central concept involved in relating contextualized examinations to the CEFR is validity (North et al., 2010), including (1) internal validity, which is the quality of the test in its own right, comprising of content validity, theory-based validity, and scoring validity (Weir, 2005); (2) external validity or concurrent validity, which adds value to the test by relating it to an external framework such as the CEFR or by using expert judgments on a CEFR panel, either inside or outside the testing organization (Papageorgiou, 2007). The validation framework, as a basis, should be embedded from the starting point of test development to later stages of test administration or test data analyses in order to provide both a theoretical perspective and a practical process for generating validity evidence.

There has been a plethora of empirical studies concerning the establishment of this type of linking argument in recent years. Martyniuk's (2010) book gathered a series of studies that looked into linking a single test to the CEFR as well as linking a suite of exams to the CEFR. The majority of these studies undertake the systematic stages of familiarization, specification, standardization and empirical validation as recommended by the Council of Europe (2009). For example, Kantarcioglu et al. (2010) reported on a study linking the Certificate of Proficiency in English (COPE) to the CEFR B2 level. The study closely followed the guidelines described in the preliminary pilot version of the publication usually referred to as the "Manual" (Council of Europe, 2003) and all four interrelated stages were undertaken. The study involved familiarization with activities suggested in the Manual, supplemented

by a series of in-house quizzes. They used a graphical profile in the specification stage and the examinee-paper selection method was employed for the writing paper. The Angoff and the YES/NO method were used for the reading and listening papers in the standardization stage, to establish the reliability of the cut-off score. Evidence was gathered from the live exam, teacher judgments, and a correlation study that was used to establish the validity of COPE. In addition, O'Sullivan's (2010) study provided empirical evidence that aimed to confirm the link between a single test, the City and Guilds' Communicator, and the CEFR B2 level by following the stages of familiarization, specification, standardization, and empirical validation.

Other researchers also used similar procedures to establish the criterion-referenced validity of the test(s) they are examining and wishing to link to the CEFR. For example, Downey et al. (2010) attempted to link the Hellenic American University's Advanced Level Certificate in English examination (ALCE) to the CEFR. The project executed the first three stages proposed in the Manual and designed an empirical validation stage for future research. The project involved an in-depth familiarization with the content and levels of the CEFR, followed by the mapping of the ALCE examination to the categories and levels of the CEFR using a graphic profile. The results of the study suggested that the ALCE test is targeted at the C1 level of the CEFR. Similarly, Khalifa et al. (2010) applied the familiarization and specification procedures to confirm the alignment of the First Certificate in English (FCE) with the CEFR B2 level. This study demonstrated that the Manual's methodology can be constructively utilized for the development of a linking argument and also for the maintenance of the alignment of a test to the CEFR.

In addition, Kecker et al. (2010) studied all four English skills, that is, writing, reading, listening, speaking, of the Test of German as a Foreign Language (TestDaF), following the steps described in the Manual to examine the relation between the TestDaF and the relevant CEFR level. The receptive skills (reading and listening) adopted a modified Angoff approach, whereas the productive skills (writing and speaking) used teachers' judgments on test performance by TestDaF candidates. Wu et al. (2010) attempted to establish an alignment of the reporting levels of the General English Proficiency Test (GEPT) reading comprehension test to the CEFR levels by following the internal validation procedures, including familiarization, specification, and standardization. The project was undertaken to meet the Taiwanese Ministry of Education requirement (issued in 2005) that all major language tests needed to be mapped to the CEFR.

The majority of the studies reviewed above are concerned with establishing *a posteriori* validity evidence for the existing test(s) by linking to the relevant CEFR levels. Not much effort has been made to collect *a priori* validity evidence alongside the test development procedures. This chapter reports on how CEFR scales were operationalized in practice in the course of developing the Pearson Test of English Academic (PTE Academic™).

PTE Academic™ is a computer-based international English language test launched globally in 2009. The purpose of the test is to assess English language competence in the context of academic programs of study where English is the language of instruction. It is targeted at intermediate to advanced English language learners. In order to claim that PTE Academic™ is fit for purpose, validity evidence has been collected during the various stages of test development through to its administration.

The constructs measured in PTE Academic™ are the communicative language skills needed for reception, production, and interaction in both oral and written modes, as these skills are considered necessary to successfully follow courses and to actively participate in the targeted tertiary level education environment.

The CEFR describes the skills language learners need to acquire in order to use a language for communication and effective action. Language ability is described within the CEFR with reference to a number of scales, which include a global scale, skill specific communicative competency scales, and linguistic competency scales. In the context of PTE Academic™, measures to link the test to the CEFR have been studied at different stages of test development. *A priori* measures include activities that incorporate the use of CEFR scales in item writing. *A posteriori* evidence includes the statistical validation procedures used to establish the extent to which PTE Academic™ scores can be linked to the CEFR.

A Priori Validation

Since test scores of PTE Academic™ are used for university admission purposes, the high-stakes nature of the decisions requires this test to be valid for the inferences the test users make, that is, whether test takers have adequate English proficiency to participate in English-medium tertiary settings. In developing valid test items, quality assurance measures were adopted at each stage of the test development process.

Qualified item writers are trained to become familiar with two essential test development documents, that is, the Test Specification (hereafter the Specification) and the Item Writer Guidelines (hereafter, the Guidelines). The Specification serves as an operational definition of the constructs the test intends to assess. The Guidelines include detailed test specification of PTE Academic™, reproduce the relevant CEFR scales, specify in detail the characteristics of each item and give item writers rules and checklists to ensure that the test items they develop are fit for purpose and suitable for inclusion in the item bank.

In developing reading and listening items, item writers are trained in three aspects: 1) familiarization with the Target Language Use (TLU) situations; 2) selection of appropriate reading or listening texts; and 3) familiarization with the CEFR scales on reading and listening. The Guidelines explain the characteristics of reading and listening passages through which test takers can best demonstrate their abilities. For the reading items, this includes text sources, authenticity, discourse type, topic, domain, text length, and cultural suitability. For the listening items, they include text sources, authenticity, discourse type, domain, topic, text length, accent, text speed, how often the material will be played, text difficulty, and cultural suitability.

In developing writing and speaking items, the Guidelines explain TLU situations with details of the CEFR scale from levels B1 to C2. In the Guidelines for writing, the purpose of writing discourse and the cognitive process of academic writing are presented in a matrix format with recommendations for preferred item types. The purposes of writing tasks are defined as 1) to reproduce, 2) to organize or reorganize, and 3) to invent or generate ideas. Three types of cognitive processing are differentiated: to learn, to inform, and to convince or persuade. In the Guidelines for speaking, item writers are instructed to produce topics focusing on academic interests

and university student life. A list of primary speaking abilities is also provided, including the ability to comprehend information and to deliver such information orally, and the ability to interact with ease in different situations.

Writing to the CEFR Levels

This section describes specific procedures involved in the writing of test items according to the CEFR levels and Table 4.1 presents an overview of the four main stages in the CEFR familiarization training for item writers. Item writers are

TABLE 4.1 *Item Writer CEFR Training Stages*

Stages	**Details**
STAGE 1: Familiarization with the definitions of some basic terms used in CEFR	For example: general language competence, communicative language competence, context, conditions and constraints, language activities, language processes, texts, themes, domains, strategies, tasks
STAGE 2: Familiarization with the common reference level: the global descriptors	**Proficient user (C2 and C1):** precision and ease with the language, naturalness, use of idiomatic expressions and colloquialisms, language used fluently and almost effortlessly, little obvious searching for expression, smoothly flowing, well-structured language **Independent user (B2 and B1):** effective argument, holding one's own, awareness of errors, correcting oneself, maintains interaction and gets across intended meaning, copes flexibly with problems in everyday life **Basic user (A2 and A1):** interacts socially, simple transactions in shops, etc., skills uneven, interacts in a simple way
STAGE 3: Familiarization with the subscales for four skills	CEFR Overall Spoken Production and subscales CEFR Overall Spoken Interaction and subscales CEFR Overall Listening Comprehension and subscales CEFR Overall Reading Comprehension and subscales CEFR Overall Written Interaction and subscales CEFR Overall Written Production and subscales
STAGE 4: Rating candidates' performances Rating communicative tasks on CEFR scales	Rate individually Express reasons and discuss with colleagues Compare rating with experts' marks

instructed to write items with a targeted difficulty level from B1 to C2 on the CEFR scale. The CEFR estimate is one of the item dimensions that item writers need to provide when they submit an item, among other item dimensions such as item source, accent and speech rate for speaking items, prompts for writing items, and distractor options for multiple choice items. Item writers' CEFR estimates of item difficulty levels were empirically validated when the items were analyzed, either through field testing or through a live item seeding process.

As shown in Table 4.1, there are four stages in the item writers' CEFR familiarization training. The first stage covers the instruction of some key terms that are used in the CEFR descriptors, aiming to facilitate item-writer trainees to understand the CEFR in general. By introducing the global descriptors at each level, the second stage gives trainees an idea of what kind of tasks and how well the test takers are expected to perform at the targeted levels. The third stage provides the trainees with more detailed descriptions of the can-do statements. Finally, after becoming familiar with the CEFR scales, item writers are asked to rate several example performances and communicative tasks individually, discuss their ratings with their colleagues, and compare their scores and reasons with those given by experts. Table 4.2 shows an example of CEFR overall written production and subscales that were used in the item-writer training.

TABLE 4.2 *An Example of CEFR Overall Written Production and Subscales*

	CEFR Overall Written Production	**CEFR Writing Subscales**
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure that helps the reader find significant points.	Creative writing Reports and essays Overall written interaction
C1	Can write clear, well-structured texts on complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples and rounding off with an appropriate conclusion.	Correspondence Notes, messages, and forms Note taking Processing text
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizng and evaluating information and arguments from a number of sources.	Orthographical control Thematic development Coherence and cohesion General linguistic range
B1	Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements in a linear sequence.	Vocabulary range Vocabulary control Coherence
A2	Can write a series of simple phrases and sentences linked with simple connectors like *and*, *but*, and *because*.	Propositional precision
A1	Can write simple isolated phrases and sentences.	

In summary, in the context of PTE Academic™, the concepts and approach of the CEFR is built into the essential test development documentation, that is, the Specification and the Guidelines. It is then implemented at the initial stage of item writing. Item writers provide a CEFR estimate for each item, which is then cross-validated at the *a posteriori* stage using statistical evidence.

A Posteriori Validation

This section reports on the statistical validation procedures used to establish the alignment of PTE Academic™ scores to the CEFR scales. Statistical procedures for relating PTE Academic™ scores to the levels of the CEFR scales involved both a test-taker-centered approach and an item-centered approach.

Linking to the CEFR: A Test-Taker—Centered Approach

For the test-taker–centered approach, data in an incomplete, overlapping design containing responses from 3,318 test takers on close to 100 items representing three item types was used. Five responses were available per test taker: Written essay (one item), Oral description of an image (two items) and Oral summary of a lecture (two items). The essay writing task has eleven score categories (0–10 points), the oral description of an image has eight score categories (0–7 points), and the oral summary of a lecture has five score categories (0–4 points), adding up to a total score of twenty-three score categories. Each response was rated on the relevant CEFR scales for writing and speaking by two human raters, independently of the ratings produced to score the test. A total of forty raters were randomly assigned to items, each rater rating on average 500 responses. Given the probabilistic and continuous nature of the CEFR scale, ratings at adjacent levels were expected in the model.

The relation between ability estimates based on scored responses on the above PTE Academic™ test items and the CEFR is displayed in Figure 4.1, with one chart for the written responses and the other for the oral responses. The horizontal axis ranges from CEFR levels A2 to C2. The vertical axis shows the truncated PTE Academic™ theta scale within the range from –2 to +2. The box plots show substantial overlap across adjacent CEFR categories as well as an apparent ceiling effect at C2 for writing. CEFR levels, however, are not to be interpreted as mutually exclusive categories. Language development is continuous and does not take place in stages. Therefore, the CEFR scale and its levels should be interpreted as probabilistic: learners of a language are estimated most likely to be at a particular level, but this does not reduce to zero their probability of being at an adjacent level.

The overlap between the box plots in Figure 4.1 is therefore in agreement with the model. Although the official CEFR literature does not provide information on the minimum probability required to be at a CEFR level, the original scaling of the levels (North, 2000) is based on the Rasch model, where cut-offs are defined at 0.5 probability. A mathematical feature that is often overlooked in aligning tests to the CEFR is therefore that the cut-offs for the ability of test takers is necessarily

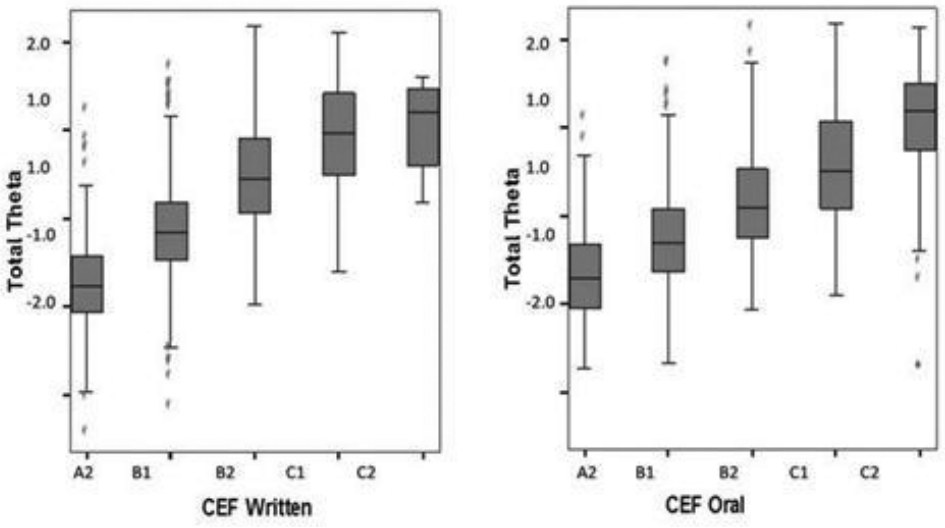

FIGURE 4.1 *CEFR Level Distribution Box Plots.*

at the midpoint between the lower and upper boundary of the level for the task difficulties. This feature is also described by Adams et al. (2002, pp. 197–199) for the level definitions in PISA (Programme for International Students Assessment). The distance of approximately 1 or 2 logits between the CEFR levels implies that anyone typically reaching a probability of around 0.8 at level X has 0.5 probability of being at level X+1, and is therefore exiting level X and entering level X+1. Having a probability of 0.5 of being at level X implies a probability of 0.15 to be at level X+1 and as little as 0.05 at level X+2.

This probabilistic relation between task difficulty and test taker performance can also be seen in Figure 4.2. A deterministic level definition where any performance is either at one level or at another would result in a step-like function. It is clear from this figure, however, that the PTE Academic™ theta increases monotonically from A2 to C2. Based on this monotone increase, a positive relation between the CEFR scale and the PTE Academic™ scale is established. To find the exact cut-offs on the PTE Academic™ theta scale corresponding to the CEFR levels, that is, the position where a particular CEFR rating becomes more likely than a rating below, the first stage is to establish the lower bounds of the CEFR categories based on the independent CEFR ratings. For this purpose, the CEFR ratings were scaled using FACETS (Linacre, 1988; 2005). The estimates of category boundaries on the CEFR theta scale are shown in Table 4.3.

The relationship between the scale underlying the CEFR levels and the PTE Academic™ theta for those test takers about whom we had information on both scales (n = 3,318) is shown in Figure 4.2. The horizontal axis shows the CEFR theta, and the vertical axis shows the PTE Academic™ theta estimate. The correlation between the two measures is 0.69. A better fitting regression is obtained with a first order polynomial (unbroken curved line), yielding an r^2 of slightly over 0.5. This

TABLE 4.3 *Category Lower Bounds on CEFR Theta*

Category	CEFR Level	CEFR Theta (Lower bounds)
1	A2	–4.24
2	B1	–1.53
3	B2	0.63
4	C1	2.07
5	C2	3.07

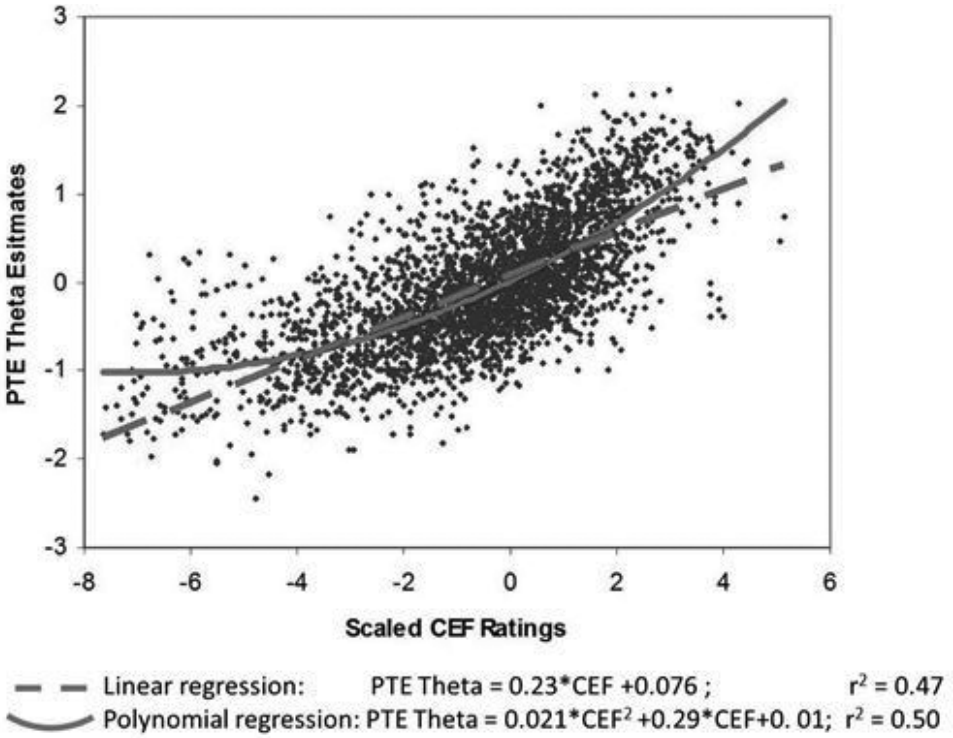

FIGURE 4.2 *Relation between CEFR Theta and PTE Theta.*

regression function was used to project the CEFR cut-offs from the CEFR scaled ratings onto the PTE Academic™ theta scale.

Because of noisy (messy and unpredictable) data at the bottom end of the scales, the lowest performing 50 candidates were removed. Further analyses were conducted with the remaining 3,268 subjects. Figure 4.3 shows the cumulative frequencies for these 3,268 candidates for whom theta estimates are available on both scales

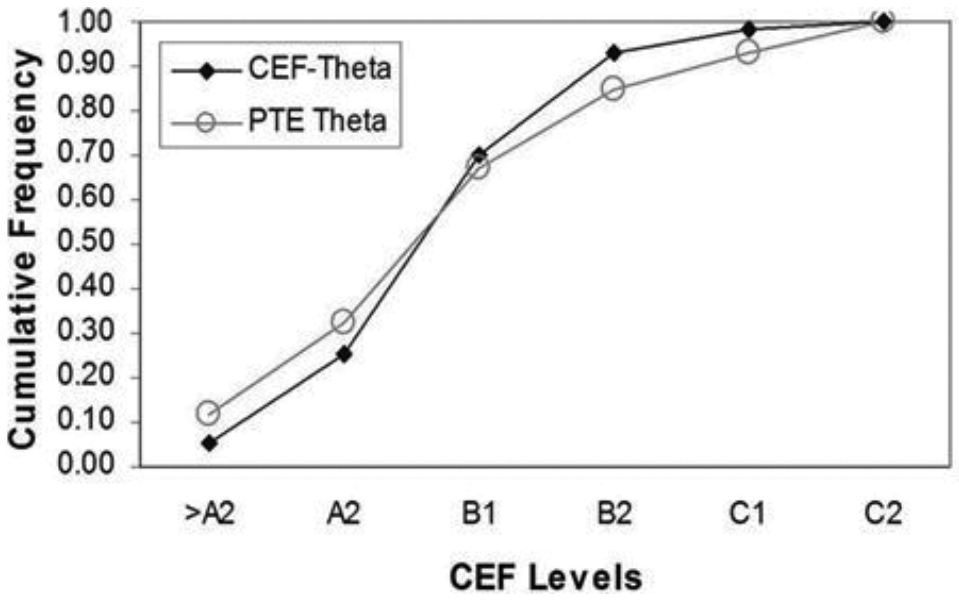

FIGURE 4.3 *Cumulative Frequencies for CEFR Levels on CEFR and PTE Theta Scales.*

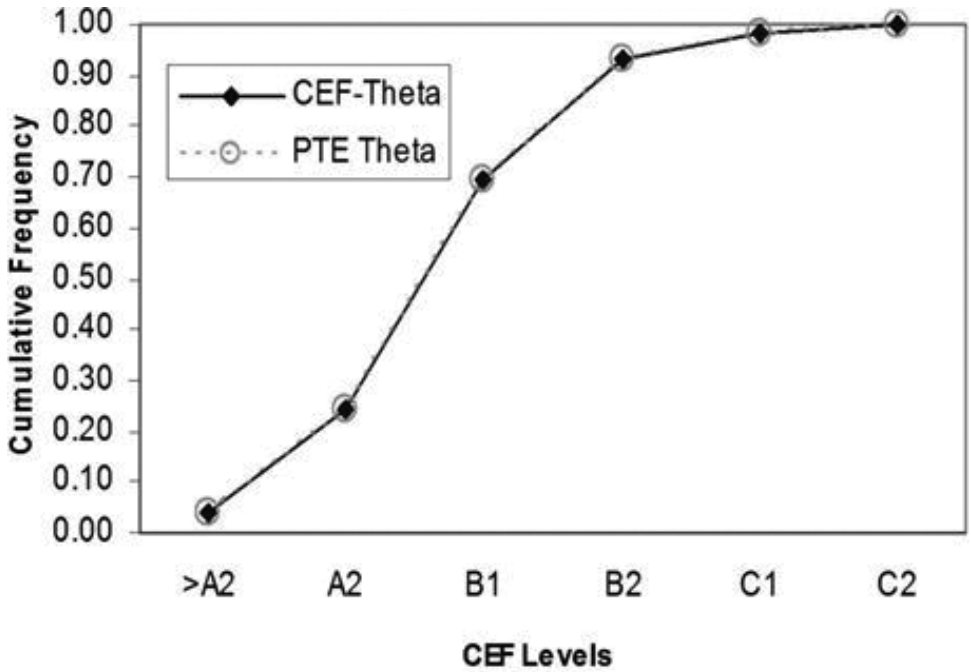

FIGURE 4.4 *Cumulative Frequencies on CEFR and PTE Theta Scales after Equipercentile Equating.*

TABLE 4.4 *Final Estimates for CEFR Lower Bounds on the PTE Theta Scale*

CEFR Levels	Theta PTE	Frequency	Percentage	Cumulative Frequency
A2	–1.155	677	21%	0.25
B1	–0.496	1,471	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	3,268	100%	

(CEFR scale and PTE Academic™ scale). The cumulative frequencies are closely aligned, though the PTE scale shows slightly less variance.

In the next stage, equipercentile equating was chosen to express the CEFR lower bounds on the PTE theta scale. Equipercentile equating of tests determines the equating relationship as one where a score has an equivalent percentile on either test. The cumulative frequencies are shown in Figure 4.4 showing a complete alignment on both scales and the resulting projection of the CEFR lower bounds on the PTE theta scale together with the observed distribution of field test candidates over the CEFR levels is shown in Table 4.4.

Linking to the CEFR: An Item-centered Approach

As reported above, at the item development stage, item writers were required to indicate, for each item, the level of ability expressed in terms of the CEFR levels they intended to measure, that is, whether they thought test takers would need to be able to correctly solve the items. In the item review process, these initial estimates from item writers were evaluated, and if needed, corrected by the item reviewers. Based on observations from field tests, the average item difficulty was calculated for items to fall into a particular category according to item writers. Table 4.5 provides the mean observed difficulty for each of the CEFR levels targeted by the item writers.

However, rather than the average difficulty of CEFR levels, the cut-offs between these levels as they are projected on the PTE Academic™ theta scale need to be established. To this effect, from the data, given item difficulty, the likelihood of any item having been assigned to any of the CEFR levels was estimated. The cut-offs between the two consecutive levels is the location on the scale where the likelihood of belonging to the first category becomes less than the likelihood of belonging to the next category. In this way, the PTE theta cut-offs based on the items were found. The estimated lower bounds of the difficulty of items targeted at each of the CEFR levels were plotted against the lower bounds of these levels as estimated from the independent CEFR ratings of test takers' responses by human raters. In Figure 4.5,

TABLE 4.5 *Intended and Observed Item Difficulty*

Intended CEFR Level	Mean Observed Difficulty
A2	0.172
B1	0.368
B2	0.823
C1	1.039
C2	1.323

the horizontal axis represents the theta scale of the CEFR cut-offs from the test-taker–centered analysis, while the vertical axis represents the scale of the CEFR cut-offs from the item-centered analysis. Note that because the analyses were conducted independently, each scale has its own origin and measurement unit. Both estimates,

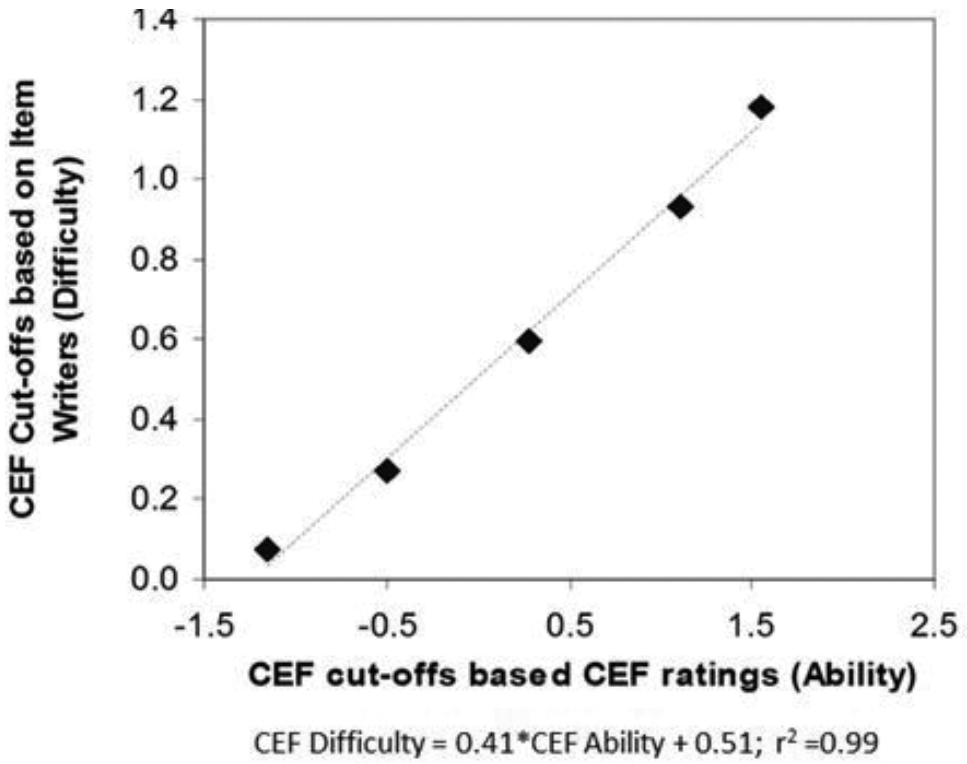

FIGURE 4.5 *Lower Bounds of CEFR Levels Based on Targeted Item Difficulty versus Lower Bounds Based on Equated CEFR Ratings of Candidates' Responses.*

derived independently, agree to a high degree (r = 0.99) on the relative distances between the cut-offs.

Provisions for the maintenance of the relationship with the CEFR need also to be made continuously as new test items are written by item writers in adherence with both Test Specification and Item Writer Guidelines. These new items are systematically seeded in the operational test forms to gather live test-taker responses. When enough responses are collected, these items are scored and analyzed, together with other precalibrated test items. The analysis adopts a concurrent design of calibration, whereby new test items, following analysis of the results, are benchmarked to CEFR-referenced item difficulties.

Concordance with Other Measures of English Language Competencies

In order to triangulate the estimated relationship with the CEFR, concordance studies were conducted between PTE Academic™ and other measures of English language competencies claiming alignment to the CEFR during the field and beta testing stages. Test takers' self-reported scores on other tests of English, including TOEIC®, TOEFL® PBT, TOEFL® CBT, TOEFL iBT®, and IELTS™ taken within two months prior to or after their participation in the PTE Academic™ field tests. In addition, test takers were asked to send in a copy of their score reports from these tests. About one in four of all test takers that provided self-reported scores also sent in their official report. The correlation between the self-reported results and the official score reports was .82 for TOEFL iBT® and .89 for IELTS™. This finding is in agreement with earlier research on self-reported data. For example, Cassady (2001) found students' self-reported GPA scores to be "remarkably similar" to official records. The data are also consistent. When the TOEFL iBT® was first released, the Educational Testing Service prepared a document (ETS, 2005) with concordances between the TOEFL iBT® and its predecessors the TOEFL® CBT and the TOEFL® PBT. This helped users of previous versions of the TOEFL to transition to the Internet-based version. In this, now historical, document (ETS, 2005, p. 7), the score range 75–95 on TOEFL iBT® is comparable to the score range 213–240 on TOEFL® CBT and to the score range 550–587 on TOEFL® PBT. Table 4.6 shows the mean of the self-reported scores in those tests and their corresponding correlation with PTE Academic™ during field testing, and Table 4.7 shows the corresponding correlation with PTE Academic™ during beta testing.

In addition, a score range of 800–850 on TOEIC® corresponds to a score range of 569–588 on TOEFL® PBT (Wilson, 2003). This is in line with data collected during the PTE Academic™ field test (see Table 4.6).

Based on the data presented in Tables 4.6 and 4.7, concordance coefficients were generated between PTE Academic™ and other tests of English using linear regression. The regression coefficients were then used to predict the scores of PTE Academic™ BETA test takers' scores on TOEFL iBT® and IELTS™. Table 4.7 shows the self-reported mean scores and those from the official reports, the mean scores from the same test takers as predicted from their PTE Academic™ score, and the correlations between the reported scores and predictions from PTE Academic™. Table 4.8 shows

TABLE 4.6 *Means and Correlations of PTE Academic™ Field Test Takers on Other Tests*

Test	Self-reported Data			Official Score Report		
	n (valid)	Mean	Correlation	n	Mean	Correlation
TOEIC®	327	831.55	.76	n/a		
TOEFL® PBT	92	572.3	.64	n/a		
TOEFL® CBT	107	240.5	.46	n/a		
TOEFL iBT®	140	92.9	.75	19	92.1	.95
IELTS™	2432	6.49	.76	169	6.61	.73

TABLE 4.7 *Correlation and Prediction of PTE Academic™ BETA Test Takers*

Test	Self-reported Data				Official Score Report			
	n	Mean	Predicted	Correlation	n	Mean	Predicted	Correlation
TOEFL iBT®	42	98.9	97.3	.75	13	92.2	98.2	.77
IELTS™	57	6.80	6.75	.73	15	6.60	6.51	.83

TABLE 4.8 *TOEFL iBT® CEFR Cut-offs Estimated by ETS*

CEFR	TOEFL iBT®	Alignment to CEFR
	Estimated by ETS	Estimated from PTE Academic™
C1	110	110–111
B2	87	87–88
B1	57	57–58

two independent approaches to estimating CEFR cut-scores for the TOEFL iBT®. One column shows the cut-scores arrived at through research conducted by ETS (Tannenbaum et al., 2008). The other column shows the estimated cut-scores based on the PTE Academic™ concordance with TOEFL iBT®. The high level of agreement between these two approaches provides independent support for the validity of the CEFR cut-offs on the PTE Academic™ reporting scale as presented in this study.

Discussion: Advantages and New Developments

Studies that link tests to the CEFR have relevance beyond supporting evidence for test validity. Once sufficiently convincing evidence for the alignment has been established, it can also help stakeholders to understand the meaning of test scores in a more comprehensive way because test scores can be interpreted in terms of the "can do" statements in the CEFR. In other words, both teachers and learners can gain better understanding of what students with a particular score are likely to be able to accomplish in terms of the descriptive system of the CEFR and its level descriptors. Such understanding can help learners self-assess their learning and assist teachers to reflect on their teaching, which could potentially make learning and teaching more effective. Future studies might also investigate the impact of the validity of the linking study from teaching and learning perspectives.

The reporting scale of PTE Academic™ uses the Global Scale of English (GSE), which is a linear scale used by Pearson to express increasing difficulty of language tasks as well as growing ability of language users. The scale runs from ten to ninety, thereby offering a more granular scale to measure progress in English language proficiency than the six levels of the CEFR can offer. Other tests can be developed that report on the same scale. In order to understand how this is possible, we need to remind the reader of the creation process of CEFR common reference levels. The six levels of the CEFR are described by a set of holistic paragraphs presented in Table 4.1 of the CEFR (Council of Europe, 2001, p. 24). These paragraphs as well as those presented in Tables 4.2 (Council of Europe 2001, pp. 26–27) and Table 4.3 (Council of Europe 2001, pp. 28–29) in fact constitute summaries of sets of illustrative descriptors or "can do" statements (Council of Europe 2001, p. 25) estimated to describe language proficiency at six predefined intervals on an underlying latent language proficiency variable. The locations of the descriptors on the underlying continuous variable were estimated in a research project reported by North (2000) and summarized in Appendix B of the CEFR (Council of Europe, 2001, pp. 217–225). After calibrating the bank of close to 500 descriptors, North (2000) divided the continuous variable in intervals which later became the CEFR levels.

Going back to the illustrative descriptors allows for the subdivision of the CEFR levels and for making the development of functional language competence reportable with greater precision. The finer resolution of the granular scale is defined by the descriptors in the form of "can do" statements. However, in order to realize this granularity at all levels of the CEFR, more illustrative descriptors are needed than are available in the CEFR or in North (2000). The CEFR offers about three times more descriptors over the range from A2 to B2 than on the rest of the scale and more than half of all descriptors relate to speaking. In a large scale project at Pearson, hundreds of new descriptors are being developed and scaled using the original North (2000) descriptors as anchors. The descriptors (grouped as "General Adult," "Academic," "Professional." and "Young Learners") together describe the Global Scale of English. The learning objectives for adults are already available free of charge (GSE; see: http://www.english.com/blog/gse-learning-objectives-for-adults/)

and versions tailored for learners of professional English, academic English, and young learners will be made available in due course.

The possibility of defining a universal scale of functional language ability across first and second and/or foreign languages (De Jong, 1984, 1988) with validity across various school systems (De Jong, 1986) and across various first language backgrounds (De Jong et al., 1990) had already been suggested and supported by a number of research projects compiled in De Jong (1991). The requirements for the psychometric definition of levels that compartmentalize a latent continuum were suggested in the context of reporting the PISA 2000 results by De Jong in an e-mail to Ray Adams and described by Adams et al. (2002).

Like the CEFR, the GSE can be used to create syllabuses, course material, and examinations, but at a more granular level, so as to make progress observable within a school year or even a semester. Further studies are needed to chart the learning time required to make progress on the scale, depending on parameters such as distance of first language from English, exposure to English and intrinsic motivation. Furthermore, studies can be conducted on the relative efficacy of learning and teaching methods. On the one hand, such studies could deepen our understanding of the relation between language proficiency levels and the actual teaching practices in the classroom. On the other hand, these studies could help to develop realistically attainable curricular standards and national language policies.

Conclusion

Linking a test to a common scale, such as the CEFR, presents its merits alongside its challenges. Establishing concurrent validity entails establishing the degree to which results from a test agree with the results from other measures of the same or similar constructs. One caveat, however, with this type of validity evidence, as Moller (1982) reminds us, is that we need to check whether or not the criterion measure itself is valid. If it is not valid or not designed to measure the same construct then one cannot claim that a test has criterion-related validity because it correlates highly with another test or external criterion of performance.

To make these kinds of linking efforts, the processes of linking should involve intertwined stages from test development to statistical validation. This chapter has reported on the measures taken to support the concurrent validity of PTE Academic™, established from the beginning of the test development process by writing items according to CEFR criteria, gaining statistical evidence to demonstrate alignment with the CEFR, and by comparing results from other tests of a similar nature that have claimed alignment with the CEFR.

The establishment of a valid link to the CEFR helps facilitate the interpretation of test scores to worldwide test users and potentially across tests of similar nature. This link should be supported by both qualitative and quantitative evidence. The linking study reported in this chapter combines *a priori* validation and *a posteriori* validation. To further consolidate validity evidence, it would be advisable to adopt a variety of other approaches to collect more qualitative data, for instance, collecting introspective justification, retrospective group discussion, and think-aloud protocols or systematic recordings.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Byram, M., & d Parmenter, L. (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Bristol, UK: Multilingual Matters.

Cassady, J. C. (2001). Self-reported GPA and SAT scores. *ERIC Digest*. ERIC Identifier: ED458216.

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

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version* DGIV/EDU/LANG (2003), 5.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment: A manual*. Strasbourg, France: Author.

De Jong, J. (1984). Listening: A single trait in first and second language learning. *Toegepaste Taalwetenschap in Artikelen*, 20, 66–79 (ERIC: ED 282 412).

De Jong, J. (1986). Achievement tests and national standards. *Studies in Educational Evaluation*, 12(3), 295–304.

De Jong, J. (1988). Rating scales and listening comprehension. *Australian Review of Applied Linguistics*, 11(2), 73–87.

De Jong, J. (1991). *Defining a variable of foreign language ability: An application of item response theory* (ISBN 90-9004299-7) (Doctoral dissertation). Twente University.

De Jong, J., & Oscarson, M. (1990). Cross-national standards: A Dutch–Swedish collaborative effort in national standardized testing. *AILA Review*, 7, 62–78.

Downey, N., & Kollias, C. (2010). Mapping the Advanced Level Certificate in English (ALCE) examination onto the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press, pp. 119–130.

Educational Testing Service. (2005). *TOEFL® Internet-based test: Score comparison tables*. Princeton, NJ: Educational Testing Service.

Kantarcioglu, E., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). Benchmarking a high-stakes proficiency exam: the COPE linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press, pp. 102–118.

Kecker, G., & Eckes, T. (2010). Putting the manual to the test: The TestDaF-CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge. UK: Cambridge University Press, pp. 50–79.

Khalifa, H., French, A., & Salamoura, A. (2010). Maintaining alignment to the CEFR: The FCE case study. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press, pp. 80–101.

Linacre, J. M. (1988). *A computer program for the analysis of multi-faceted data*. Chicago, IL: Mesa Press.

Martyniuk, W. (Ed.), (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press.

Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th edn.). New York, NY: Macmillan, pp. 1487–1495.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.

Moller, A. D. (1982). *A study in the validation of proficiency tests of English as a foreign language* (Unpublished doctoral thesis). University of Edinburgh, Scotland.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York, NY: Peter Lang.

North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relating language examinations to the common European framework of reference for languages in the context of the Council of Europe's work on language education. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press, pp. 1–17.

O'Sullivan, B. (2010). The City and Guilds Communicator examination linking project: A brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual.* Cambridge, UK: Cambridge University Press, pp. 33–49.

Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference*. London, UK: Trinity College London.

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement: Interdisciplinary research and perspectives*, 2(3), 178–182.

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.* Educational Testing Service Research Report. RR-08-34, TOEFLiBT-06. Princeton, NJ: Educational Testing Service.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Oxford, UK: Palgrave.

Wilson, K. (2003). *TOEFL® Institutional Testing Program (ITP) and TOIC® Institutional Program (IP): Two on-site testing tools from ETS at a glance*. Handout from Educational Testing Service at Expolingua, Berlin, Germany.

Wu, J. R. W., & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press, pp. 204–224.