CRELLA
University of
Bedfordshire

Relating the Versant English Placement Test to the Common European Framework of Reference

Anthony Green

Centre for Research in English Language Learning and Assessment

University of Bedfordshire

May 2013

Executive Summary

A project was undertaken by the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire in collaboration with Pearson Knowledge Technologies (the test developer) to relate the Versant English Placement Test (VEPT) to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001), generally known as the CEFR. The project included both an appraisal of which aspects of the CEFR framework are addressed by the VEPT and expert judgement on how test material and test taker performance relate to the CEFR levels.

A standard setting procedure was conducted following the guidelines of the *Manual for Relating Language Examinations to the Common European Framework of Reference* (Council of Europe, 2009). The standard setting procedure began with a specification exercise to determine which aspects of the CEFR are addressed in the VEPT. This specification was reviewed individually by six consultants at the University of Bedfordshire. After a review of the specification and CEFR familiarization training, a group of expert judges was assembled to act as a panel for the purpose of linking the VEPT to the CEFR. These experts included teachers, applied linguists, researchers in language testing, and test developers. Three different standard setting approaches were used to establish the relationship between the VEPT and the CEFR: 1) the Basket method, 2) a person-centred performance rating method, and 3) the Body of Work method. For the first approach, panelists were presented with 111 items and were asked '*At what CEFR level can a test-taker already answer the following item correctly?*' This approach was applied to the following tasks in the VEPT that elicit short responses: Repeats, Sentence Builds, Conversations, Sentence Completion, and Dictation. For the second approach, panelists were presented with 108 test-taker responses and were asked *'On the evidence of this performance, at what CEFR level would you place this learner?'* This approach was applied to the following tasks in VEPT that elicit more extended responses: Read Aloud, Passage Reconstruction, and Summary and Opinion. For the third approach, panelists judged 8 test-takers' performances on the test as a whole.

Since the items presented to the panelists already had difficulty estimates and the performances had already been scored on the VEPT scale, the items and the performances on the VEPT scale could be compared to the panelists' CEFR judgments. Multi-facet Rasch analysis was used because it places item difficulty and test-taker ability on the same measurement scale and also takes into account the relative harshness or leniency of the judges. Regression analysis was then used to relate the two and

to establish what cut scores on the VEPT should be used to place learners into different CEFR levels. Because the three approaches suggested somewhat different relationships between the VEPT and the CEFR, the results from the three approaches were averaged and rounded up to the nearest integer (or whole score point).

Table 1 summarises the recommendations made by the expert panellists on the general relationship between scores on the VEPT and the CEFR levels.

Table 1. Recommended Mapping of CEFR Levels with Versant English Placement Test Overall Scores

VEPT (20-70)	CEFR (A1-C1)
20-23	<A1
24-33	A1
34-45	A2
46-56	B1
57-65	B2
66-70	C1

This table is presented with two important caveats:

The scope of the test in relation to the aspects of language ability covered by the framework must be taken into account when interpreting the alignment of scores to levels. The VEPT does not involve all of the features of language covered in the CEFR. For example, no specific claims are made in relation to sociolinguistic competences, as these are not addressed by the test. It should also be noted that the overall picture that the test scores provide of a learner’s CEFR level may mask strengths or weaknesses in specific areas. The further the interpretation of the results departs from the aspects of language directly covered by the test, the more caution must be applied.

The second caveat is that CEFR levels may be interpreted differently according to context. The definition of learner level employed by the panellists engaged in this study is not necessarily the interpretation used in every language school. Schools may therefore wish to adjust entry points for classes working towards CEFR objectives according to local requirements and based on their experience. Further research will be needed locally (within schools) and internationally (across teaching and learning contexts) to establish the extent to which these recommendations are consistent with other sources of evidence such as teacher judgements and the results of other tests.

Introduction

This report summarises a project undertaken by the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire in collaboration with Pearson Knowledge Technologies (the test developer) to relate the Versant English Placement Test (VEPT) to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001), generally known as the CEFR.

The Common European Framework of Reference for Languages

The CEFR is designed to bring together the best of a wide range of different schemes for describing language learning levels to provide a common language for professionals working in this area. The CEFR summarises the scope of a global consensus on levels of functional communicative language ability and lays out options for successful language learning, posing important questions to inform and guide the development of locally appropriate resources for teaching, learning and assessment.

In the process of developing levels for the CEFR, *Can Do* statements were gathered from 30 different schemes, giving an initial pool of over 2,000 statements to be considered for inclusion (North, 2000). These were screened for repetition and then approved, rejected or edited by the CEFR authors (who reduced the number by half) to leave a set of meaningful, ‘positively worded, “stand-alone” statements’ (North, 2000, p.184). Users of the framework should be able to understand each statement without having to compare it with other statements and be able to relate the statements to learners’ practices and abilities. The framework is intended to provide a shared language for setting objectives, developing materials and making comparisons between systems. Learners should be able to look at a statement and say, ‘yes, I can do this;’ ‘no. I can’t do this;’ or ‘I’d like to learn to do this’ when using a given language.

Through having teachers rank the statements and use them to judge the performance of learners, North (2000) constructed an overarching set of common reference levels representing expanding communicative language ability. At the end of the development process, each *Can Do* statement and each performance had been ‘calibrated:’ given a mathematical value to allow estimates of the probability that a learner judged able to perform one activity would be able to perform any of the other activities appearing on the scale. The teachers involved in the development were also asked to group the statements into categories to represent different aspects of language use.

The outcome of these processes was a network of interrelated scales (Chapters 4 and 5 of the CEFR), categorised into a *Descriptive Scheme* covering a range of features that might be addressed in language education around the world. The descriptors are ordered by teachers according to a consensus view of their relative difficulty, illustrative of loosely defined underlying levels of ability or *Common Reference Levels*. In the ‘branching approach’ adopted, distinctions are made between three general levels (A: basic, B: independent and C: proficient), while the scales provided in the CEFR generally define differences at six levels: A1 and A2, B1 and B2 and C1 and C2. Finer-grained divisions are often made, as between B2.1 and B2.2 and it is suggested that yet more delicate distinctions can be made within each level so that the relatively small gains in language proficiency occurring within language programmes (achievement) can be captured and reported. The CEFR scales are not a closed system: additional sets of *Can Do* statements, such as the ALTE *Can Do* scales presented at as Appendix D of the CEFR (Council of Europe [COE], 2001, p. 244ff), can be integrated into the CEFR scheme using appropriate statistical techniques.

The CEFR *Descriptive Scheme* and the *Common Reference Levels* together provide a conceptual grid which users can utilize to describe their systems. Since its publication, the CEFR has been widely adopted as a convenient means of comparing language qualifications designed for different purposes, produced in different countries and developed by different assessment agencies. The adoption of CEFR levels by policy makers in setting targets for language learning achievement or requirements for the language abilities of migrants has encouraged testing agencies to link their qualifications to the CEFR and encouraged the Council of Europe to provide guidance on defensible methods for doing so. Such linking does not imply an equivalence between the results of different testing systems related to the CEFR at the same level (different tests are developed with different purposes in mind), but, when taken in conjunction with an awareness of the test content and coverage, should help users and others to 'locate' the test in CEFR terms.

To facilitate this process, the Council of Europe published *Relating examinations to the Common European Framework of Reference: A Manual* (henceforth referred to as 'the Manual') released in draft form in 2003 and updated in 2009.

The Manual presents three related sets of procedures that users are advised to follow in order to link assessments to the CEFR:

- Specification of examination content
- Standardisation of judgments
- Empirical validation through analysis of test data

However, the Manual stresses that it does not provide the 'sole guide to linking a test to the CEFR' (p.1). Indeed, the process of linking that it advocates has attracted some criticism, particularly where such exercises are conducted as 'one-off events' (Milanovic 2010, p.4) rather than embedded in the routine operation of an assessment. Effective linking of a testing system can only occur when test results are reliable and consistent over time and so continue to reflect the same standards in relation to the same aspects of the CEFR.

In relating the VEPT to the CEFR, the project described here broadly follows the steps advocated in the Manual, but also provides an account of how the approach to the routine production of test material ensures consistent levels of item difficulty.

The Versant English Placement Test (VEPT)

The Versant English Placement Test (VEPT) is an English proficiency test. The test is delivered automatically by computer on screen and via a microphone headset. It is intended for non-native English speaking adults over the age of 16.

As set out by the test developer, Pearson Knowledge Technologies (2012), the VEPT measures 'facility in spoken and written English' (p.3). Facility in spoken and written English is how well a person can understand spoken and written English and respond appropriately in speaking and writing on everyday topics, at a native-like pace in intelligible English. Results may be used for such decisions as placement testing, exit testing, and progress monitoring by academic and government institutions as well as commercial and business organizations.

The VEPT consists of 81 items and takes 50 minutes to administer. The test questions are designed to reflect a wide range of situations in the classroom and real world. There are nine task types: Read Aloud, Repeats, Sentence Builds, Conversations, Typing, Sentence Completion, Dictation, Passage Reconstruction, and Summary and Opinion. These tasks are designed to provide multiple, independent measures that underlie facility in spoken and written English, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, pronunciation of rhythmic and segmental units, and appropriateness and accuracy of writing. In relating the test to the CEFR, the initial step of specification involves connecting these concepts to the descriptive scheme used in the framework.

Test Format

The following subsections provide brief descriptions given by the test developers of the tasks and abilities required to respond to the items in each of the nine parts of the VEPT.

The Parts of the VEPT include:

A. Read Aloud
B. Repeats
C. Sentence Builds
D. Conversations
E. Typing
F. Sentence Completion
G. Dictation
H. Passage Reconstruction
I. Summary & Opinion

Part A: Read Aloud

In Part A, test takers read out two short passages – expository texts that deal with generally familiar everyday topics. The test takers are given 30 seconds to read each text while it is displayed on the computer screen. All passages are relatively simple in structure and vocabulary and range in length from 60 to 70 words. They have a SMOG readability score between 5 and 7 and can be read easily and fluently by most educated English speakers.

For test takers with little facility in spoken English but with some reading skills, this task provides an opportunity to display their pronunciation and oral reading fluency. In addition to information on reading rate, rhythm, and pronunciation, the scoring of the Read Aloud task is informed by miscues. Miscues occur when a reading is different from the words on the page or screen, and provide information about how well test takers can make sense of what they read. For example, hesitations or word substitutions are likely when the decoding process falters or cannot keep up with the current reading speed; word omissions are likely when meaning is impaired or interrupted. More experienced readers draw on the syntax and punctuation of the passage, as well as their knowledge of commonly co-occurring word patterns; they can monitor their rate of articulation and comprehension accordingly. This ability to monitor rate helps ensure that reading is steady as well as rhythmic, with correct stress and intonation that conveys the author’s intended meaning. Less experienced readers are less able to comprehend, articulate and monitor simultaneously, resulting in miscues and breaks in the flow of reading. The Read Aloud section appears first in the test because, for some test takers, reading aloud presents a familiar task and is a comfortable introduction to the interactive mode of the test as a whole.

Part B: Repeats

In Part B, test takers are asked to repeat verbatim the sentences that they hear. The sentences are presented to the test taker in approximate order of increasing difficulty. Sentences range in length from 3 to 15 words. The audio item prompts are spoken in a conversational manner.

To repeat a sentence longer than about seven syllables, a person must recognize the words as spoken in a continuous stream of speech (Miller & Isard, 1963). Highly proficient speakers of English can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with English words, phrase structures, and other common syntactic forms. If a person habitually processes five-word phrases as a unit (e.g. "the really big apple tree"), then that person can usually repeat utterances of 15 or 20 words in length. Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not familiar with English sentence structure.

Because the Repeat items require test takers to organize speech into linguistic units, Repeat items assess the test taker's mastery of phrase and sentence structure. Given that the task requires the test taker to repeat full sentences (as opposed to just words and phrases), it also offers a sample of the test taker's fluency and pronunciation in continuous spoken English.

Part C: Sentence Builds

For Part C, test takers hear three short phrases and are asked to rearrange them to make a sentence. The phrases are presented in a random order (excluding the original word order), and the test taker says a reasonable and grammatical sentence that comprises exactly the three given phrases.

To correctly complete this task, a test taker must understand the possible meanings of the phrases and know how they might combine with other phrasal material, both with regard to syntax and pragmatics. The length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one-word versus a three-word phrase) that a person can hold in verbal working memory. This is important to measure because it reflects the test taker's ability to access and retrieve lexical items and to build phrases and clause structures automatically. The more automatic these processes are, the more the test taker's facility in spoken English. This skill is demonstrably distinct from memory span (see Section 2.5, Test Construct, below). The Sentence Builds task involves constructing and articulating entire sentences. As such, it is a measure of test takers' mastery of sentences in addition to their pronunciation and fluency.

Part D: Conversations

In Part D, test takers listen to conversations between two speakers. Each conversation typically consists of an exchange made up of three short sentences. Immediately after each conversation, a recorded voice asks a comprehension question. Test takers answer the question with a word or short phrase.

This task measures test takers' listening comprehension ability. Conversations are recorded at a conversational pace covering a range of topics. The task requires test takers to follow speaking turns and extract the topic and content from the interaction at a conversational pace. Quick word recognition and decoding and efficient comprehension of meaning are critical in correctly answering the question.

Part E: Typing

In the Typing task, test takers see a passage at the top of the computer screen and have 60 seconds to type the passage exactly as it appears into a box at the bottom of the screen. All passages deal with general everyday topics. The passages are relatively simple in structure and vocabulary and range in length from 90 to 100 words. The SMOG Readability Index was used to identify and refine the readability score for each passage. All passages have a readability score between 7 and 8, which can be easily typed by most educated English speakers with adequate typing skills.

This task has several functions. First, it allows test takers to familiarize themselves with the keyboard. Second, it measures the test taker's typing speed and accuracy. The Versant English Placement Test assumes a basic competence in typing for every test taker. Since it is important to disambiguate test takers' typing skills from their written English proficiency, it is recommended that test administrators review each test taker's typing score. If typing speed is below 12 Words per Minute, and/or accuracy is below 90%, then it is likely that this test taker's written English proficiency was not properly measured due to poor typing skills. The test administrator should take this into account when interpreting test scores.

Part F: Sentence Completion

In this task, test takers read a sentence that has a word missing, and they supply an appropriate word to complete the sentence. Test takers are given 25 seconds for each item. During this time, test takers must read and understand the sentence, retrieve a lexical item to complete the sentence, and type the word in the text box provided. Sentences range in length from 5 to 25 words. Across all items in this task, test takers are exposed to sentences with words missing from various parts of speech (e.g., noun, verb, adjective, adverb) and from different positions in sentences: sentence-initial, sentence-medial, sentence-final.

It is sometimes thought that fill-in-the-gap tasks (also called cloze tasks) are more authentic when longer passages or paragraphs are presented to the test taker, as this enables context-inference strategies. However, research has shown that test takers rarely need to look beyond the immediate sentence in order to infer the correct word to fill the gap (Sigott, 2004). This is the case even when test designers specifically design items to ensure that test takers go beyond sentence-level information (Storey, 1997). Readers commonly rely on sentence-level comprehension strategies partly because the sentence surrounding the gap provides clues about the missing word's part of speech and morphology and partly because sentences are the most common units for transmission of written communication and usually contain sufficient context for meaning. Above and beyond knowledge of grammar and semantics, the task requires knowledge of word use and collocation as they occur in natural language. For example, in the sentence: "The police set up a road ____ to prevent the robbers from escaping," some grammatical and semantically correct words that might fit include "obstacle", "blockage" or "impediment." However, these would seem inappropriate word choices to a native reader, whose familiarity with word sequences in English would lead them to expect a word such as “block” or “blockade.” In many Sentence Completion items there is more than one possible correct answer choice. However, all items have been piloted with native speakers and learners of English and have been carefully reviewed with reference to content, collocation and syntax. The precise nature of each item and possible answer choices are quantified in the scoring models. The Sentence Completion task draws on interpretation, inference, lexical selection and morphological encoding, and as such reflects the test taker’s mastery of vocabulary in use.

Part G: Dictation

In Part G, test takers hear a recorded sentence and must type the sentence exactly as they hear it. Test takers have 25 seconds to type each sentence. The sentences are presented in approximate order of increasing difficulty. Sentences range in length from 3 words to 14 words. The items present a range of grammatical and syntactic structures, including imperatives, wh-questions, contractions, plurals, possessives, various tenses, and particles. The audio item prompts are spoken with a natural pace and rhythm by various native and non-native speaker voices that are distinct from the examiner voice.

Dictation requires the test taker to perform time-constrained processing of the meanings of words in sentence context. The task is conceived as a test of expectancy grammar (Oller, 1971). An expectancy grammar is a system that governs the use of a language for someone who has knowledge of that language. Proficient listeners tend to understand and remember the content of a message, and not the exact words used; they retain the message rather than the words that carry the message. Therefore, when writing down what they have heard, test takers need to use their knowledge of the language either to retain the word string in short term memory or to reconstruct the sentence that they have forgotten. Those with good knowledge of English words, phrase structures, and other common syntactic forms can keep their attention focused on meaning, and fill in the words or morphemes that they did not attend to directly in order to reconstruct the text accurately (Buck, 2001, p. 78). The task is a good test of comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for test takers who are not familiar with English words and sentence structures. Analysis of errors made during dictation reveals that the errors relate not only to interpretation of the acoustic signal and phonemic identification, but also to communicative and productive skills such as syntax and morphology (Oakeshott-Taylor, 1977).

Part H: Passage Reconstruction

Passage Reconstruction is similar to a task known as free-recall, or immediate-recall. Test takers are required to read a text, put it aside, and then write what they can remember from the text. In this task, a short passage is presented for 30 seconds, after which the passage disappears and the test taker has 90 seconds to reconstruct the content of the passage in writing. Passages range in length from 45 to 65 words. The items sample a range of sentence lengths, syntactic variation and complexity. The passages are short stories about common situations involving characters, actions, events, reasons, consequences, or results. In order to accurately reconstruct a passage, the test taker must read the passage presented, understand the concepts and details, and hold them in short term memory in order to reconstruct the passage. Individual test takers may naturally employ different strategies when performing the task. Reconstruction may be somewhat verbatim in some cases, especially for shorter passages answered by advanced test takers. For longer texts, reconstruction may be accomplished by paraphrasing and drawing on the test taker’s own choice of words. Regardless of strategy, the end result is evaluated based on the test taker's ability to reproduce the key points and details of the source passage using grammatical and appropriate writing. The task requires the kinds of skills and core language competencies that are necessary for activities such as responding to requests in writing, replying to emails, recording events or decisions, or summarizing texts.

The Passage Reconstruction task is held to be a purer measure of reading comprehension than, for example, multiple-choice reading comprehension questions, because test questions do not intervene between the reader and the passage. It is thought that when the passage is reconstructed in the test taker's mother tongue then the main ability assessed is reading comprehension, but when the passage is reconstructed in the target language (in this case, English), then it is more an integrated test of both reading and writing (Alderson, 2000, p. 230).

Part I: Summary and Opinion

In Part I, test takers are presented with a passage. They are given 18 minutes to read the passage, write a summary of the author's opinion in 25 to 50 words, and give their own opinion on the topic presented in the passage in at least 50 words. The passages contain an opinion on an everyday topic. All passages consist of an introduction, two body paragraphs, and a conclusion. All passages are relatively simple in structure, use vocabulary from the most frequently-occurring 1,200 words in English, and range in length from 275 to 300 words. The SMOG Readability Index was used to identify and refine the readability score for each passage. All passages have a readability score around 10, which are easily understandable by most educated English speakers.

In the Summary response, test takers are expected to demonstrate a clear understanding of the author's opinion and to identify the supporting points without including unnecessary details or repeated information. In order to do so, the test taker must read the passage presented, understand the concepts and details, and evaluate the information presented in order to identify the most important points. Responses are scored on the quality of the summary and writing conventions. In the Opinion response, test takers are expected to provide their own opinion on the topic presented and to provide clear and appropriate supporting ideas and/or examples. Test takers must construct an informative response with appropriate spelling, punctuation, capitalization, syntax, and grammar. Responses are scored on the quality of the opinion and writing conventions. The Summary and Opinion task draws on reading comprehension, interpretation, inference, summarization, syntax, and writing mechanics, and as such reflects the test taker's mastery of reading and writing.

Scoring

Versant English Placement Test results are reported in the form of an Overall score and four skill scores (Speaking, Listening, Reading, and Writing).

Overall Score

The Overall score of the test represents the ability to understand spoken and written English and respond appropriately in speaking and writing on everyday topics, at a native-like pace and in intelligible English.

The overall score is a weighted combination of the four skill scores and is reported on a scale ranging from 20 to 70.

Previous work (Pearson Education, 2011) had suggested the following correspondences between VEPT Overall scores and CEFR levels (Table 2):

Table 2. Mapping of CEFR Levels with Versant English Placement Test Overall Scores (Pearson Education 2011, p.8)

VEPT (20-70)	CEFR (A1-C1)
20-23	<A1
23-32	A1
33-45	A2
46-55	B1
56-67	B2
68-70	C1

Speaking

Speaking reflects the ability to produce English phrases and clauses in complete sentences. The score is based on the ability to produce consonants, vowels, and stress in a native-like manner, use accurate syntactic processing and appropriate usage of words in meaningful sentence structures, as well as use appropriate rhythm, phrasing, and timing.

Listening

Listening reflects the ability to understand specific details and main ideas from everyday English speech. The score is based on the ability to track meaning and infer the message from English that is spoken at a conversational pace.

Reading

Reading reflects the ability to understand written English texts on everyday topics. The score is based on the ability to operate at functional speeds to extract meaning, infer the message, and respond appropriately.

Writing

Writing reflects the ability to produce written English texts on everyday topics. The score is based on the ability to present ideas and information in a clear and logical sequence, use a wide range of appropriate words as well as a variety of sentences structures.

Of the 81 items on the VEPT, 75 responses are currently used in the automatic scoring. The first item response in Parts B, C, D, F, and G are considered practice items and are not incorporated into the final score.

Figure 1 illustrates which tasks of the test contribute to each of the four skill scores. Each vertical rectangle represents a response from a test taker. The items that are not included in the automatic scoring are shown in blue.

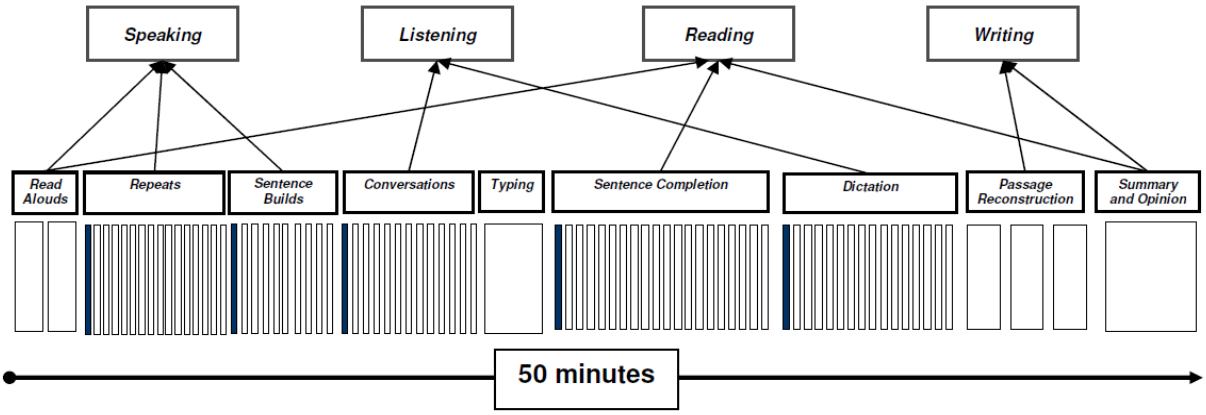

Figure 1. Relation of skill scores to tasks (from Pearson 2012, p. 16)

Relating the VEPT to the CEFR

The main focus of this report is on the process of standard setting involving panels of expert judges. A description of the methods employed is preceded by a brief outline of the specification phase.

Specification

Following initial familiarisation with the CEFR, as recommended in the Manual (Council of Europe 2009), staff at Pearson Knowledge Technologies carried out a *Specification* exercise to arrive at initial estimates of the CEFR-VEPT relationship. This involved identifying common ground between the VEPT and the CEFR and using forms from Section A2 of the Manual (*Forms for Describing the Examination*) to identify which aspects of the CEFR are addressed by the test, whether directly or indirectly and how the conceptualisation of language abilities underlying the test connects to their conceptualisation in the CEFR descriptive scheme. Evidence from the CEFR and from previous standard setting studies was included in support of the decisions made.

The initial specification was reviewed individually by six consultants at the University of Bedfordshire, who subsequently came together in a half-day workshop to respond to this initial draft. The consultants considered how well the specification claims appeared to be supported by evidence from the CEFR and made detailed comments on the forms. Pearson Knowledge Technologies then revised their claims in the light of this feedback to arrive at the version included here as Appendix 2.

In addition to informing preliminary hypotheses about the VEPT-CEFR relationship in terms of language levels and the range of abilities addressed, the process of specification served to highlight a number of issues for the consultants. They pointed in particular to the potential discrepancy between the functionally oriented CEFR descriptors and the nature of the VEPT tasks. For example, at the B1 level, the general characterisation provided in the CEFR suggests that learners 'Can deal with most situations likely to arise whilst travelling in an area where the language is spoken' (Council of Europe, 2001:24). The consultants observed that the VEPT tasks do not generally involve realistic communication in defined settings, but generally focus instead on the manipulation of language forms. The VEPT tasks do not directly simulate the language activities described in the CEFR, but seem to focus more on syntax, lexis and phonology.

This focus on linguistic form in the VEPT is consistent with long established practice in language testing. Tests in this tradition have demonstrated value, particularly as placement instruments, and there would appear to be a close relationship between performance on such tests and communicative language ability. This is reflected in the CEFR which locates linguistic competences on the same overarching measurement scale as communicative language activities and strategies. Nonetheless, as Hulstijn (2007) has observed, there is little empirical evidence of the degree of linguistic competence as conceived in the CEFR (lexical competence; grammatical competence; semantic competence; phonological competence; orthographic competence; orthoepic competence) that would enable a language learner to function in the ways described in the illustrative scales, nor are key task conditions – such as the amount of planning time available or the support available from interlocutors – consistently specified in the framework (Green 2012). In other words, the consultants suggested, relating performance on VEPT tasks to functional CEFR descriptors might involve more inference on the part of panellists than would be the case for tests that do simulate language in use in social settings.

The process of specification also indicated to the consultants that it would be inappropriate to adopt a single approach to standard setting when relating the VEPT to the CEFR. One reason for this is the complexity of the VEPT and of its scoring algorithm. Dichotomously scored items and performance tests scored using rating scales are each generally associated with different approaches to standard setting. The VEPT includes a combination of item types with a variety of response formats (see above). While the automated scoring system rates each response to each VEPT item on a scale, some sections (such as Repeats and Sentence Completion) would seem to have more the character of dichotomous (right/wrong) items while others (such as Summary and Opinion) seemed to be closer to traditional scaled items.

It is assumed in the CEFR Manual (and this is reflected in the forms used for Standardisation) that each part of a test will contribute uniquely to one component score. The sum of the scores on the Reading sub-test will give a Reading score, a Writing sub-test gives a Writing score. However, in the VEPT (see Figure 1) each test Part may contribute to more than one component score. Read Aloud items contribute to scores for both Reading and Speaking while Summary and Opinion items contribute to scores for Reading and Writing. This feature of the VEPT added to the complexity of the linking process.

Standard Setting

In the Council of Europe Manual, it is stated that,

> ‘It is possible that multiple standards have to be set for the same test. In linking to the CEFR, one might wish, for example, to set a cut score for A2, B1 and B2. It is important to understand what is precisely meant by the preceding sentence. A cut score is to be conceived as a border between two adjacent categories on some scale. So the example should be understood in the sense that every test taker will be classified either as A2, B1 or B2, and hence we need two cut scores: one that marks the border between A2 and B1 and one for the border between B1 and B2. In general the number of cut scores is one less than the number of classification categories’ (Council of Europe 2009: 70).

Naturally, a placement test like the VEPT will require multiple cut-scores as learners need to be placed into classes at a range of levels. Multiple cut scores would therefore need to be identified.

Determining the appropriate cut-scores hinges on the identification of what is rather insensitively termed the 'minimally acceptable person' at each CEFR level. The concept is explained in the Manual as follows:

> 'A basic concept, which also appears in many other standard setting procedures, is the concept of the "minimally acceptable person", also referred to sometimes as the "borderline person" or person "just barely passing" or "minimally competent test taker". Where a standard has to be set, for example, for CEFR Level B1, a minimally acceptable person has the competencies, skills and abilities to be labelled as "B1", but only to such an extent that the slightest decrease in those competencies, skills and abilities would suffice in order not to grant this qualification. The task for the panellists is to keep in mind such a person or collection of persons during all the judgmental work they have to do' (Council of Europe 2009, p.73).

Although it is considered basic to any linking process, in the Manual the meaning of '[having] the competencies, skills and abilities to be labelled as "B1"'; is unfortunately ambiguous. In the CEFR each level represents clusters of language use activities that have a ('vertical') range of difficulty and a ('horizontal') spread of functional purpose (viewed from three perspectives: competences, activities and strategies). The Manual does not specify how much or what balance of these attributes would qualify a test taker for a label of B1. In language learning programmes (the context for the VEPT), a label of 'B1 learner' might typically be applied to one who is working towards selected B1 activities as learning objectives. In proficiency testing on the other hand, a 'B1 test taker' would more usually be one who had already demonstrated achievement of B1 objectives. The two approaches within testing practice that have most often been taken include a 'mastery' approach – in which a test taker needs to satisfy all of the relevant criteria in order to be awarded a certain score – and a 'best fit' approach – in which awarding a score would involve identifying the level on a scale that seems to best describe the test taker's abilities (even where the criteria associated with the level are not all fully satisfied). A B1 test taker would be one whose competences, skills and abilities seem better captured by the general B1 level description than by A2 or B2.

Therefore, in linking to the CEFR there are at least three possible interpretations of the 'minimally competent B1 learner'. First, this could refer to a learner who is working towards objectives that fall on the B1 side of the A2-B1 border on the scale of descriptors (even if the learner may be unable to perform most B1 level activities). Second, it might refer to a learner who is (just) better described by the overall characterisation of the B1 level than the A2 level. Third, it could describe a learner who satisfactorily demonstrates the ability to perform all B1 level activities that are relevant to a specific teaching/ learning context.

For the purpose of linking the VEPT to the CEFR, the 'best fit' approach appears the most defensible. Performance on the VEPT does not provide direct evidence of how test takers might perform on a specific activity and so seems more compatible with the broad characterisations of proficiency offered by the common reference levels (CEFR Section 3.6) than with any particular illustrative scale. Additionally, as the CEFR is an under-specified and open-ended system, it is not possible to determine whether or not a learner has ever ‘mastered’ a given level of the framework. Scales and descriptors are simply not available for many areas of the descriptive scheme (which is not in any case restricted to the categories presented in the framework document). The ‘mastery’ interpretation is unlikely to be consistent with the organisation of classes in the contexts for VEPT as a learner who has mastered all relevant B1 objectives should not be placed into a B1 level class, but into a B2 level class or higher.

Both the VEPT and the CEFR recognise that individual learners’ abilities are not always well captured by a single overall score. The Manual suggests that it is possible to offer ‘profiles’ of both test takers and tests in relation to the framework categories – indeed such a profile is the outcome of the suggested specification stage of the linking process (see below). In the CEFR, a good deal of emphasis is given to the partial, variable and uneven nature of language competence and it is recognised that a learner may be stronger in certain areas of linguistic and functional competence than others – more proficient in spoken than written language or better able to function in the personal than public domain, for example. The VEPT reports scores for each of the traditional ‘four skills’ of Reading, Writing, Listening and Speaking (but not following the CEFR scheme of production, interaction, reception and mediation). On the other hand, for placement purposes a single overall estimate of ability is usually required.

All VEPT responses are scored automatically by computer. However, automated scoring systems are generally developed by first obtaining large numbers of ratings made by human judges and then ‘training’ a computer system to emulate the human raters. In the case of the VEPT, a set of eight separate rating scales is used by human judges to score test taker performances. To explore the relationship between the VEPT and the CEFR, these eight scales (oral reading fluency, fluency, reading pronunciation, pronunciation, narrative clarity and accuracy, summary writing, opinion writing, writing conventions) were compared by panellists with the 57 scales presented in the CEFR. Similarities of wording were used to build up a picture of the relationship between the two.

For dichotomous items, the Manual suggests a variation on the modified Angoff standard setting technique known as the Basket Method. Panellists are presented with a set of test items and are asked to respond to the question, ‘*At what CEFR level can a test taker already answer the following item correctly?*’ The total number of items on a test that a panellist would expect a minimally competent test taker at B1 level to answer correctly is taken as that panellist’s recommendation (in raw score terms) for the B1 cut-score. Results for all panellists are aggregated and, following one or more rounds of discussion, recommended cut-scores are arrived at that represent the view of the panel as a whole.

Although there is in fact no dichotomous scoring of items on the VEPT, it was suggested by Pearson Knowledge Technologies that, for the purposes of linking, items that involved short answers could be treated as though they were scored dichotomously. This meant that a form of Basket Method could be applied for the following Parts of the VEPT:

- Repeats
- Sentence Builds
- Conversations
- Sentence Completion
- Dictation

The Manual suggests an Extended Tucker-Angoff Method for items that are scored on a scale. In this method, the question posed to panellists is:

> *'Suppose that 100 borderline persons answer the item, where one can earn up to [4] points, what would be in your view the average score obtained by these 100 persons?'*

However, as the scores awarded would depend on the interpretation of the VEPT scales it was considered that applying this method might simply repeat the comparison of CEFR and VEPT scales already undertaken. Furthermore, a test taker centred approach using the CEFR scales directly to rate performance on test tasks was believed to offer a more intuitive and straightforward task for the panellists. Where sets of extended samples were available, these were rated against the CEFR levels. The test Parts for which this approach was adopted included:

- Read Aloud
- Passage Reconstruction
- Summary and Opinion

In addition, a variation of the Body of Work Method was used to judge the overall performance of a small number of test takers on each test Part. Panellists read or listened to the responses of each test taker to one Part of the test and awarded a score against the CEFR General Reference Levels.

Given the complexity of the VEPT scoring algorithm, cut scores for the test as a whole could not be calculated by simply adding results for the individual items. However, both item difficulty estimates and person ability measures for VEPT are reported on a common scale. It was therefore possible to establish cut scores for the test as a whole expressed on a common metric using results from the different standard setting approaches.

The CEFR linking panel

As recommended in the Manual, a group of expert judges was assembled to act as a panel for the purpose of linking the VEPT to the CEFR. These experts included teachers, applied linguists, researchers in language testing and members of the team responsible for developing the tests. The following sections describe the participants and procedures followed.

Participants

Two separate panels were established for the linking process. Both were made up of experienced and well-qualified English Language educators and both followed the same procedures. The first panel was made up of seven independent teachers and researchers either working for the University of Bedfordshire, or recruited as qualified experts for the event. The second panel was made up of seven employees of Pearson Knowledge Technologies, the developer of the VEPT.

Including both independent experts and VEPT 'insiders' in this way was intended to ensure that the judgements both of those with a close knowledge of the test and of those with a more independent perspective would be included in the linking process. Although it would have been preferable to include both groups in a single panel, this was not possible for logistical reasons including the physical distance between the University of Bedfordshire (Luton, UK) and Pearson Knowledge Technologies (Menlo Park, California). The difference in time zones also ruled out the option of holding a joint panel via video conferencing.

The UK panellists

Four of these panellists were teachers with extensive experience of teaching English as a foreign language at a range of levels. Three were researchers familiar with the CEFR working in the field of applied linguists or language testing and working for the University of Bedfordshire (the UK consultants to the linking project). All of the UK panellists worked in the Higher Education sector in the UK. One had less than five years' experience in English language education, two had between 11 and 15 years, one between 16 and 20 and three had more than 20 years' experience. Five of the panellists held diplomas in English Language Teaching for Adults (DELTA or RSA Dip. TEFLA), six of the seven had Masters in Applied Linguistics or ELT and three held PhDs.

All but one had previously participated in a CEFR linking panel. Six of the seven claimed a 'Good' knowledge of the CEFR; the seventh had a 'Basic' knowledge. Additional experience with the CEFR included developing tests, language programmes and rating scales based on the CEFR levels, and publications evaluating the CEFR or describing how tests are related to the CEFR.

The US panellists

The seven US panellists were employees of Pearson Knowledge Technologies who were more familiar with the VEPT test, but generally had less experience as educators and less familiarity with the CEFR. Although seven took part in the panel, two did not respond to the online familiarisation task and one did not provide background information. Four of the six US panellists who did provide background information had less than five years' experience with English language education, one had between six and ten and one between 11 and 15. Two reported a Masters level qualification (one in TESOL, the other in Linguistics). Two reported a 'Good' knowledge of the CEFR, two a 'Basic knowledge and one a 'Poor' knowledge. The sixth did not respond to this question. Only one had previously participated in a CEFR linking panel, but one had developed a rating scale based on the CEFR, two had used the CEFR to rate spoken performance and another reported that 'I reference it regularly in my test development projects'.

Familiarisation training

Prior to the panel events, all panellists were asked to familiarise (or re-familiarise) themselves with the Common European Framework levels by reviewing the illustrative scales. Specifically, they were directed a) to the general reference levels presented on pages 28-29 of the English version of the Framework; b) to Section 3.6, which provides an overview of salient features of each level; and c) to relevant illustrative scales from Chapters 4 and 5 of the CEFR. Those who were less familiar with the framework or who felt that they needed further guidance were also asked to undertake self-access training using the *CEFTrain* website (www.helsinki.fi/project/ceftrain). As a check on their interpretation, the panellists carried out an online descriptor sorting activity in which they identified the CEFR levels of thirty-two descriptors taken from areas of the Framework that were of greatest relevance.

The Panel events

Both the UK and US panels took place over one day and followed the same agenda. They opened with a series of familiarisation activities, as recommended in the Manual (Council of Europe 2009). The agenda for the UK panel is given here and the US panel followed essentially the same timetable.

Session 1: 9:00am – 10:30am

Familiarisation

Activity 1: c.30 minutes

For the purpose of group discussion, the panellists were arranged in groups. First, working individually, each panellist sorted a jumbled set of descriptors based on Table A1 of the Manual (a summary of Section of 3.6 of the CEFR) from which references to the levels had been removed. The panellists then discussed their decisions within their group, justifying or reconsidering their choices. This was followed by a plenary discussion. Table A1 was then distributed.

Activity 2: c.20 minutes

The panellists used the CEFR self-assessment grid (Table 1.2 of the 'Structured overview of all CEFR scales') to rate themselves on two languages they could use as a foreign or additional language. They justified their self-ratings in CEFR Can Do terms and raised issues that they had encountered in making their judgements. This was followed by a general discussion on issues arising from the process including consideration of variable skills profiles and the implications of different domains of language use.

Activity 3: c.40 minutes

Training in assessing performance in relation to the CEFR levels using standardised samples

The panellists viewed illustrative samples of learners speaking English from the training DVD supplied by the Council of Europe and awarded scores expressed in terms of CEFR levels to each. They then discussed and justified their scores to their group and to the panel as a whole. Finally, the panellists compared their scores and the justifications they had given with those provided by the Council of Europe.

The panellists viewed sample performances from the Council of Europe illustrating different CEFR levels. These samples can be accessed at:

http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/videos/english.php

The three learners of English viewed by the panel were 'Clara', 'Lucas' and 'Charlotte'. Each student was viewed performing in both the monologue and dialogue condition. The panellists were not shown the scores given by the Council of Europe.

After watching each student, the panellists, working individually, assigned a score. These scores were then compared and discussed in relation to the CEFR 'Qualitative aspects of spoken language use' (CEFR Table 1.3). Following discussion the group arrived at a consensus on the appropriate score and was then shown the score awarded and 'comments on the assigned levels' provided on the website.

Session 2: 11:00 – 12:45

Scale comparisons

In this session, the focus was on the rating scales. The panellists were given copies of the rating scales used by human raters in scoring Versant performances during the test's development. The ratings by human judges had been used in training the automated scoring system employed in the VEPT.

The panellists compared these Versant scales with the illustrative scales of the CEFR. On this basis, they suggested which CEFR levels best matched each overall band level on the Versant rating scales. The panellists justified and discussed their judgements and were given an opportunity to revise these before making individual recommendations on correspondences. Each pair/trio of panellists looked at the scales. They compared these scales with the 'Structured overview of all CEFR scales' (Council of Europe 2001). The panellists identified correspondences between the scales and suggested how they felt the two related to each other.

In plenary discussion, the group matched CEFR levels to scale points on the Versant scales, noting which CEFR scales seemed to be reflected in each Versant scale and which levels on each appeared to describe similar performance features.

Session 3: 1:45pm – 3:30pm

Rating sample performances

The panellists listened to or read sample responses to VEPT items from the 'Read Aloud', 'Passage Reconstruction', and 'Summary and Opinion' sections. They responded individually to the following question:

On the evidence of this performance, at what CEFR level would you place this learner?

The outcome was a set of grids with a CEFR score from each panellist for each sample response. The panellists did not see the VEPT scores for each response, but had opportunities to discuss their ratings following each round of judgements.

Session 4: 3:45pm – 5:45pm

The panel listened to and read the responses for six test takers with overall scores on the VEPT ranging from 22 to 69 (although they were not told these scores). They (individually) assigned a CEFR level to each test taker on each section of the test. i.e. a panellist might rate one test taker as A0 (below A1) on 'Read Aloud', but A2 on 'Sentence Completion' and A1+ on 'Passage Reconstruction' etc. Again they did not see the Versant scores originally awarded.

At the end of the day, the panellists completed a short questionnaire on the event and the decisions reached.

Analysis

The objective of the panel events was to obtain judgements of the difficulty of VEPT material and the ability of VEPT test takers in relation to the CEFR. As the performances presented to the panels had already been scored and the difficulty of the items estimated on the VEPT measurement scale, the panellists' judgements could then be used as a basis for linking VEPT scores to CEFR levels.

Because it allows the difficulty of test material and the ability of test takers to be located on the same measurement scale and because it takes account of the relative harshness or leniency of the judges involved, multi-facet Rasch analysis was used in scaling the descriptors employed in the CEFR scales. The advantages of this technique suggested that it would be equally useful in analysing the outcomes of the panel events.

Once estimates of the CEFR level of test material or performances had been obtained, these could be compared to VEPT item difficulty estimates or test takers' VEPT scores. Regression analysis was then used to relate the two and to establish what cut scores on the VEPT should be used to place learners into different CEFR levels.

Results

Specification

Table 3. Form A8: Initial Estimation of Overall Examination Level

Initial Estimation of Overall CEFR Level		
☒ A1	☒ **B1**	☒**C1**
☒	☒	□
☒**A2**	☒ **B2**	□ **C2**
☒	☒	□

Short rationale, reference to documentation

Items at the A1 and A2 level were written specifically to match CEFR “Can do” statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

For more evidence see:

Versant English Test Validation Report (p. 21-23 describe CEFR concordance)
Versant English Placement Technical Paper
Bernstein, Van Moere, & Cheng (2010)

The initial estimate of the examination level, showing how the test is intended to span levels A1 to C1 is shown in Table 3 above (Form A8 of the Manual). The test developer completed Forms A1 to A21 and A24 of the Manual and sent their conclusions to the consultants for comment. Six consultants from Bedfordshire then completed a VEPT test, reviewed the specification tables and supporting evidence and commented on these.

There was general agreement that the test would cover levels A1 to C1, but it was observed that the nature of the test construct made it challenging to draw direct connections between test content and the functional activities specified in the CEFR illustrative scales. This became increasingly problematic at higher levels of the CEFR as the VEPT does not provide direct evidence of the sociolinguistic and pragmatic aspects of competence that come increasingly to the fore at the higher levels.

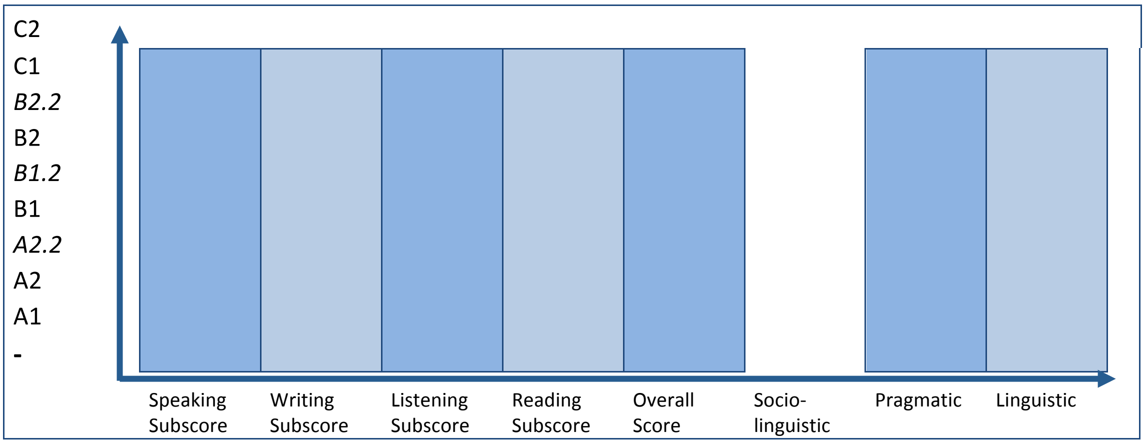

Figure 2. Form A23 of the Manual: Graphic Profile of the Relationship of the Examination to CEFR Levels

The conclusion arrived at by the test developer following the feedback from the consultants is displayed in the form of the 'graphic profile' recommended in the Manual (Figure 2, Form A23). On the basis of the feedback, the decision was taken to exclude the category of 'sociolinguistic competence' from the profile as this is not directly evidenced in the test.

Forms A9 to A21 and A24 of the Manual, illustrating the process of specification, listing sources of evidence and outlining the conclusion are given in Appendix 2.

Familiarisation

A few days before the panel events, panellists were invited to participate individually in an online survey to gauge their familiarity with the CEFR levels. They were asked to assign a CEFR level to each of 36 descriptor statements taken from the CEFR or from the related DIALANG project reported in Alderson (2005). The responses to the survey for two of the US panellists were not registered, but the results for the twelve remaining panellists are reported in Table 4.

Table 4. Number of descriptors located at each level: CEFR and panellist judgements

	Level assigned by majority of panellists					
CEFR Level	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
A1	**2**	2				
A2		**7**	1			
B1		4	**4**			
B2			5	**5**		
C1				1	**3**	
C2						**2**

Overall, the majority of panellists correctly identified the CEFR level of 23 (64%) of the 36 statements. In two cases, A1 descriptors were identified as A2 by a majority of the panellists, in one case, an A2 descriptor was placed at B1. Four B1 descriptors were placed by the majority at A2, five B2 descriptors at B1 and one C1 descriptor at B2. This suggests a degree of 'central tendency' on the part of the panellists or a tendency to cautiously locate descriptors towards the middle of the scale: overrating low level descriptors at A1 and A2 and underrating high level descriptors at B1 or above. In two cases, the group was equally split between those placing the descriptor at its CEFR level and those placing it at an adjacent level (see Appendix 1).

Of the 432 individual judgements made by the panellists, 222 (51%) were accurate: the panellist correctly identified the CEFR level of the descriptor. Perhaps reflecting their greater familiarity with the CEFR, the UK panellists were generally better able to identify the levels: 56% of their judgements were correct compared with 46% for the US panel. However, the second most accurate panellist was from the US group. Most of the incorrect judgements placed the descriptor in question one CEFR level above or below its CEFR level. The panellist able to match the highest number of descriptors correctly placed 26 (72%). The weakest performer only placed 6 (17%) at the correct level.

The online survey confirmed that members of both panels had a good understanding of the CEFR levels, although there was scope to further refine this through familiarisation activities on the day of the panel event.

Scale comparisons

Following further familiarisation activities on the day of the panel, which served to re-orient the panellists in relation to the CEFR and to build consensus, the groups discussed the VEPT scales and tried to relate these to CEFR illustrative scales and levels.

The US and UK panels approached the scale matching task rather differently and so their results are reported separately here. Briefly, US panellists attempted to match every Versant scale point to a CEFR level while the UK panellists were more selective and only drew parallels where they could evidence a connection. This meant that where no parallels suggested themselves they left more gaps than the US panellists: they identified no matches at all in the CEFR for three of the Versant scales. Second, the UK panel worked together to build one consensus view of the relationship while the US panellists worked individually and each recorded their own interpretation.

Table 5. The UK Panel: scale comparisons

Versant scale	*Points on scale*	*1*	*2*	*3*	*4*	*5*	*6*	*CEFR Illustrative scales referenced*
Fluency	6							Public announcements
Opinion Writing	4	A2+	B1	B1+	B2+	-	-	Reports and essays
Passage Reconstruction	6		A2		B1			Processing text, Overall written production
Pronunciation	6							
Reading Fluency	6							
Reading Pronunciation	6	A0	A1	A2	B1	B2	C1	Phonological control
Summary Writing	3	A2	B1	B2	-	-		Processing text
Writing Conventions	4	A1	A2	B1	B2+	-	-	Orthographic control

The UK panel concluded that certain of the VEPT scales had no clear parallel in the CEFR (Table 5). For example, the focus on the mechanics of reading aloud in the VEPT oral reading fluency scale were not reflective of the more functional descriptions provided by the CEFR. Although both the Versant and CEFR scales include reference to pausing and hesitation, the reasons for this seemed to be rather different. The CEFR *Spoken Fluency* scale indicates that at B1 learners may experience 'problems with formulation resulting in pauses'.

In the case of the VEPT *Fluency* scale, it was noted that there were no descriptors concerned with intelligibility and no specification of how much would be understood by the listener, which are central to the CEFR scale for *Phonological Control*.

Table 6. The US Panel: scale comparisons

Versant scale	*Points on scale*	*1*	*2*	*3*	*4*	*5*	*6*	*CEFR Illustrative scales referenced*
Fluency	6	A1	A2	B1	B2	C1	C1+	Spoken fluency (Reading for orientation)
Opinion Writing	4	B1	B1	B2	C1	-	-	Reports and essays (Creative writing, Overall written production)

Passage Reconstruction	6	A1	A2	B1	B1+	B2	C1	Overall written production (Processing text, Grammatical accuracy)
Pronunciation	6	A1	A1	A2	B1	B2	C1	Phonological control
Reading Fluency	6	A1	A2	B1	B2+	C1	C2	Spoken fluency (Reading for orientation)
Reading Pronunciation	6	A0	A1	A2	B1	B2	C1	Phonological control
Summary Writing	3	A1	B1	B2	-	-	-	Reports and essays, Processing text
Writing Conventions	4	A2	B1	B2	C1	-	-	Orthographic control, Grammatical accuracy

Table 6 shows the CEFR levels most often identified by US panellists with each Versant scale point. Although most US panellists matched every VEPT scale point to a CEFR level, two identified no links for *Writing Conventions* and one identified none for *Reading Fluency.* Three panellists did not identify one or more VEPT scale points on other scales with a CEFR level: usually either at the bottom (1) or top (6) of the scale affected.

The two panels were in full agreement on *Reading Pronunciation* across the full range of levels. However, where they disagreed, the tendency was for the US panel to link VEPT scale points a little higher on the CEFR than the UK panel. For example, the two panels agreed that 2 on the VEPT *Opinion Writing* scale (which both identified with the *Reports and Essays* scale of the CEFR) could be related to B1 on the CEFR. However, they disagreed on the other levels with the US panel identifying 3 and 4 with B2 and C1 respectively while the UK panel placed them at B1+ and B2+. On *Summary Writing* they agreed that 2 and 3 related to B1 and B2, but disagreed on 1 with the UK panel identifying it with A1 and the US panel A2. The US panel also interpreted the *Writing Conventions* scale as being a little higher in relation to the CEFR (A2 to C1 compared to A2 to B2+ for the UK panel). On *Passage Reconstruction*, VEPT 2 was CEFR A2 and 4 was B1 for the UK panel, but they were B1 and B2 for the US panel.

These small differences in interpretation might have been resolved through discussion had the two panels been able to work together. On the other hand, it is worth noting that standard setting is dependent on the interpretations made by panellists and diversity of opinion is to be expected and the use of separate panels reminds us of this important fact. If we had relied on either the UK panel alone or the US panel alone, the outcomes of the project would have been a little different.

Basket Method: Repeats, Sentence Builds, Conversations, Sentence Completion, Dictation

Items from these Parts of the test were treated as dichotomous (right/wrong) for the purposes of linking and were assigned a CEFR level by the panel using the Basket Method. Reflecting the CEFR, panellists were allowed to assign '+' levels such as A2+ or B1+ as well as A0 (below the scope of the CEFR) and the six CEFR levels (A1, A2, B1, B2, C1, C2). This gave a scale with a total of 14 points ranging from A0 to C2+. The ratings were converted into numeric scores for the purpose of analysis so that they could be used in making calculations:

A0	=	1
A0+	=	2
A1	=	3
A1+	=	4
A2	=	5
A2+	=	6
B1	=	7

B1+	=	8
B2	=	9
B2+	=	10
C1	=	11
C1+	=	12
C2	=	13
C2+	=	14

Table 7. Sentence Build and Repeat Item Ratings: All Facet Vertical Summary

```
+-----------------------------------------+
 |Measr|+Raters              |+Item |CEFR |
 |-----+---------------------+------+-----|
 |   2 +                     +      +(C1) |
 |     |                     |      |     |
 |     |                     |      | B2+ |
 |     |                     |      |     |
 |     |                     | *    |     |
 |     | UK 04               | *    | --- |
 |     |                     |      |     |
 |   1 +                     +      +     |
 |     | UK 07               | *    |  B2 |
 |     | UK 03               |      |     |
 |     |                     | ***  |     |
 |     |                     | **   | --- |
 |     | UK 06  US AA        |      |     |
 |     | UK 02  US GG        | *    | B1+ |
 *   0 *                     * **** *     *
 |     | US DD               |      |     |
 |     | UK 05  US CC  US EE | ***  | --- |
 |     |                     |      |     |
 |     |                     | **   |  B1 |
 |     | US BB               | **   |     |
 |     |                     | **   |     |
 |  -1 + UK 01  US FF        + *    + --- |
 |     |                     |      |     |
 |     |                     | **   | A2+ |
 |     |                     |      |     |
 |     |                     | *    |     |
 |     |                     | *    | --- |
 |     |                     | *    |     |
 |  -2 +                     +      +  A2 |
 |     |                     | **   |     |
 |     |                     | ***  |     |
 |     |                     | *    | --- |
 |     |                     | *    |     |
 |     |                     | *    |     |
 |     |                     |      | A1+ |
 |  -3 +                     +      +     |
 |     |                     |      |     |
 |     |                     | **   |     |
 |     |                     | *    | --- |
 |     |                     |      |     |
 |     |                     |      |     |
 |     |                     |      |     |
 |  -4 +                     +      +     |
 |     |                     | *    |     |
 |     |                     |      |     |
 |     |                     |      |     |
 |     |                     |      |     |
 |     |                     |      |     |
 |     |                     |      |     |
 |  -5 +                     +      + (A1)|
 |-----+---------------------+------+-----|
 |Measr|+Raters              | * = 1|CEFR |
 +----------------------------------------+
```

Table 7 is a visual representation or 'ruler' that shows the interpretation of the CEFR by the panel members (or 'raters'), placing the raters and the estimated difficulty of the test items on a common metric. The first column (Measure) shows the measurement scale derived from the analysis centred on 0 and expressed in units called 'logits'. The final column, headed 'Scale', shows this metric expressed as CEFR levels. The 'Items' in the third column, each represented by an asterisk, are the VEPT items whose difficulty was estimated by the panellists. Items towards the top of the scale were judged to be more difficult than those shown at the lower end.

The majority of the 40 Repeat and Sentence Build items were judged to range in difficulty from A1+ to B2. This is consistent with the picture of the effective range of the test that emerges from the specification and scale comparisons and suggests that these Parts of the test discriminate between learners most effectively at the B1/B2 borderline. Although relatively few items appear at A1 and A2, this does not imply that Repeat and Sentence Build items are incapable of discriminating between learners at these levels. Rather it suggests that less than 50% of learners at the A1 and A2 levels will be able to give a correct response to most of the items.

The second column ('Raters') represents the relative severity of the panellists. The US panellists are labelled from USAA to USGG and the UK panellists from UK01 to UK 07. Raters who judged the material as more difficult appear towards the top of the column, those who rated it as easy appear closer to the bottom. In contrast to the broad range of values for the items, the raters or panellists all fall within a span of 2.24 logits or, in CEFR terms (as indicated by the third column), from the A2+/B1+ boundary up to B2. The results for raters are expanded in Table 8 below.

Table 8. Rater Measurement Report for Sentence Build and Repeat items (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
UK 01	210	40	5.25	5.03	-1.02	0.15	1.63	2.35	1.37	0.99	0.76
UK 02	264	40	6.6	6.54	0.08	0.14	0.72	-1.35	0.7	-1.35	0.92
UK 03	294	40	7.35	7.35	0.66	0.14	1.08	0.43	1.05	0.3	0.77
UK 04	324	40	8.1	8.16	1.22	0.14	0.75	-1.19	0.78	-1.01	0.84
UK 05	248	40	6.2	6.09	-0.23	0.14	0.98	0	0.91	-0.31	0.85
UK 06	274	40	6.85	6.81	0.27	0.14	1.15	0.7	1.23	1.02	0.79
UK 07	306	40	7.65	7.68	0.89	0.14	1.02	0.16	1.02	0.16	0.79
US AA	275	40	6.88	6.84	0.29	0.14	1.15	0.74	1.11	0.56	0.78
US BB	224	40	5.6	5.42	-0.72	0.14	0.87	-0.51	0.87	-0.31	0.83
US CC	243	40	6.08	5.95	-0.33	0.14	0.6	-2.05	0.65	-1.49	0.86
US DD	255	40	6.38	6.29	-0.09	0.14	0.86	-0.57	0.82	-0.7	0.83
US EE	248	40	6.2	6.09	-0.23	0.14	1.11	0.55	1.61	2.16	0.74
US FF	212	40	5.3	5.08	-0.98	0.15	1.33	1.38	1.16	0.53	0.73
US GG	270	40	6.75	6.7	0.2	0.14	0.7	-1.45	0.74	-1.16	0.94

In Table 8, the Column labelled 'Observed Average' shows the average of the ratings made by each panellist. The 'Fair Average' and 'Measure' columns show estimates of the severity or harshness of each rater's judgements. It is estimated that the harshest rater, UK01, would tend to judge an item of average difficulty as 5.03 (A2) while the most lenient rater, UK04 would estimate the same item at 8.16 (B1+).

The fit statistics in the columns headed ‘InfitMS’ and ‘OutfitMS’ (standardised Infit and Outfit Mean Squares) show the degree of randomness – unexpected ratings – in the measurements. Mean-Squares usually average 1.0. Values greater than 1.0 indicate misfit or unpredictability: raters awarding scores less consistently than expected, suggesting problems such as idiosyncracies in the interpretation of the scales. Values below 1 indicate overfit or greater consistency than expected – as when raters collude in their judgements or are conservative in restricting their ratings to part of the scale. It is suggested in the FACETS user manual that figures in the range of 0.5 to 1.5 can be considered as ‘productive for measurement’ and that figures over 2.0 are potentially distorting. Although Table 8 reveals that the panellists did differ significantly (p<.01) in their degree of harshness, the figures for the Mean Squares (ranging from USCC at 0.60 to UK01 at 1.63) suggest that all were acceptably consistent in their judgements without being overly conservative.

The column headed ‘PtBis’ shows the single rater – rest of raters (SR-RoR) correlation: the correlation between the ratings made by each individual panellist and the ratings awarded by the other panellists. Myford and Wolfe (2003) suggest that SR-RoR correlations greater than .70 are high for rating scale data while correlations of .30 should be considered low. In this case, all SR-RoR correlations were greater than .70, indicating good levels of agreement between panellists in the ranking of items.

Table 9. Item Measurement Report for Repeat and Sentence Build items (arranged by estimated item difficulty)

Item	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
Repeat 2 143	120	14	8.57	8.61	0.53	0.23	1.23	0.71	1.21	0.65
Repeat 2 360	114	14	8.14	8.17	0.21	0.23	1.89	2.04	1.89	2.05
Repeat 2 510	110	14	7.86	7.87	0	0.23	0.6	-1.14	0.59	-1.19
Repeat 2 1087	66	14	4.71	4.68	-2.31	0.24	1.35	0.99	1.27	0.82
Repeat 2 1138	64	14	4.57	4.53	-2.43	0.25	0.89	-0.18	0.84	-0.34
Repeat 2 1174	79	14	5.64	5.62	-1.59	0.23	0.83	-0.38	0.86	-0.28
Repeat 2 1314	52	14	3.71	3.58	-3.28	0.3	0.79	-0.42	0.88	-0.02
Repeat 2 1407	134	14	9.57	9.55	1.26	0.25	1.4	1.1	1.51	1.33
Repeat 2 1433	45	14	3.21	3.13	-4.17	0.46	1.23	0.54	2.01	1.03
Repeat 2 1450	117	14	8.36	8.39	0.37	0.23	0.73	-0.69	0.72	-0.71
Repeat 2 2416	99	14	7.07	7.07	-0.56	0.23	0.83	-0.36	0.82	-0.38
Repeat 2 3100	50	14	3.57	3.43	-3.47	0.32	1.46	1.07	1.1	0.37
Repeat 2 3108	109	14	7.79	7.8	-0.05	0.23	1.05	0.25	1.05	0.26
Repeat 2 3229	110	14	7.86	7.87	0	0.23	0.74	-0.66	0.72	-0.71
Repeat 2 3235	66	14	4.71	4.68	-2.31	0.24	0.67	-0.92	0.66	-0.98
Repeat 2 3248	61	14	4.36	4.3	-2.61	0.25	0.71	-0.79	0.66	-0.92
Repeat 2 3349	70	14	5	4.98	-2.08	0.24	0.92	-0.1	0.92	-0.11
Repeat 2 3350	117	14	8.36	8.39	0.37	0.23	0.87	-0.26	0.87	-0.23
Repeat 2 3351	100	14	7.14	7.14	-0.51	0.23	0.47	-1.65	0.47	-1.66
Repeat 2 3358	90	14	6.43	6.42	-1.02	0.23	0.93	-0.08	0.93	-0.06
Repeat 2 3359	110	14	7.86	7.87	0	0.23	0.51	-1.54	0.51	-1.5
Repeat 2 3363	105	14	7.5	7.5	-0.25	0.23	0.71	-0.74	0.73	-0.69
Repeat 2 3364	76	14	5.43	5.4	-1.75	0.23	0.95	-0.01	0.99	0.09
Repeat 2 3367	103	14	7.36	7.36	-0.35	0.23	2.27	2.66	2.27	2.66
Sentence build 20 252	75	14	5.36	5.33	-1.8	0.23	0.61	-1.12	0.61	-1.12
Sentence build 20 496	126	14	9	9.04	0.86	0.24	0.91	-0.11	0.95	-0.01
Sentence build 20 763	103	14	7.36	7.36	-0.35	0.23	0.92	-0.11	0.91	-0.12
Sentence build 20 860	52	14	3.71	3.58	-3.28	0.3	0.75	-0.55	0.56	-0.73

Sentence build 20 880	84	14	6	5.98	-1.32	0.23	0.55	-1.34	0.54	-1.39
Sentence build 20 976	94	14	6.71	6.71	-0.81	0.23	1.35	0.96	1.32	0.9
Sentence build 20 979	93	14	6.64	6.64	-0.86	0.23	0.84	-0.33	0.84	-0.32
Sentence build 20 1001	121	14	8.64	8.68	0.58	0.23	1.57	1.43	1.56	1.41
Sentence build 20 1522	96	14	6.86	6.85	-0.71	0.23	0.42	-1.89	0.41	-1.91
Sentence build 20 1742	86	14	6.14	6.13	-1.22	0.23	1.13	0.47	1.16	0.54
Sentence build 20 1752	66	14	4.71	4.68	-2.31	0.24	1.8	1.91	1.83	1.98
Sentence build 20 1869	60	14	4.29	4.22	-2.68	0.26	0.91	-0.15	0.91	-0.11
Sentence build 20 1871	135	14	9.64	9.7	1.39	0.25	0.46	-1.76	0.43	-1.85
Sentence build 20 1873	97	14	6.93	6.93	-0.66	0.23	1.5	1.29	1.52	1.31
Sentence build 20 1882	70	14	5	4.98	-2.08	0.24	0.82	-0.39	0.84	-0.34

Table 9 shows how these items are located on the measurement scale. The relationship between persons (test takers) and items can be interpreted in terms of probabilities. A test taker with a B1 level of ability (7.0 on this numeric scale) would be estimated to have a 50% chance of success on an item rated as 7.0 such as *Repeat 2 2416*. In other words, in a group of 100 marginally competent B1 learners, the panellists would expect 50 to respond correctly to this item.

In this case, a few of the items fall outside the range that is considered productive for measurement and are either unproductive (>1.5 or <0.5), or, in one case (*Repeat 2 3367*), potentially distorting. Although it is suggested that such distorting items may be removed for the purpose of analysis, the facts that only a single item is involved and that it was only marginally outside the acceptable range suggest that this is unnecessary: all difficulty estimates were retained in the subsequent analysis.

Table 10. Conversation Item Ratings: All Facet Vertical Summary

```
+---------------------------------------+
|Measr|+Raters               |+Item|CEFR |
|-----+---------------------+-----+-----|
|   2 +                     +     +(B2+)|
|     |                     |     | --- |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |  B2 |
|     |                     |     |     |
|     |                     | *   |     |
|     | UK 07               | *   |     |
|     | UK 04               | *   | --- |
|   1 +                     + *   +     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     | *   | B1+ |
|     |                     | *   |     |
|     | UK 03  US CC        | *   |     |
|     | US BB  US DD  US EE |     | --- |
|     |                     | *   |     |
|     |                     |     |     |
|     |                     |     |     |
*   0 * UK 06               * *   *  B1 *
|     |                     |     |     |
|     |                     |     |     |
|     | US AA               |     |     |
|     | US GG               | *   | --- |
|     | UK 02               |     |     |
|     |                     | **  |     |
|     |                     |     |     |
|     |                     | *   | A2+ |
|     |                     |     |     |
|  -1 + UK 05               +     +     |
|     | UK 01  US FF        | **  | --- |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |  A2 |
|     |                     |     |     |
|     |                     |     |     |
|     |                     | *   |     |
|     |                     |     | --- |
|  -2 +                     + *   +     |
|     |                     |     |     |
|     |                     | *   |     |
|     |                     |     | A1+ |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     |     |
|     |                     |     | --- |
|  -3 +                     +     + (A1)|
|-----+---------------------+-----+-----|
|Measr|+Raters               | * = |CEFR |
|-----+---------------------+-----+-----|
```

Table 10 shows that the 18 Conversation items were judged to range in difficulty from A1+ to B1+ with a broad distribution covering these CEFR levels. There was a relatively wide distribution of panellist judgements on Conversation items of 3.45 logits or two CEFR levels separating the harshest from the most lenient panellist.

Table 11. Rater Measurement Report for Conversation items (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
US AA	113	18	6.28	6.26	-0.35	0.21	0.68	-1.03	0.68	-1.05	0.78
US BB	130	18	7.22	7.21	0.37	0.21	1.18	0.63	1.17	0.6	0.58
US CC	134	18	7.44	7.43	0.54	0.21	2.04	2.51	2.06	2.55	0.45
US DD	130	18	7.22	7.21	0.37	0.21	1.19	0.66	1.2	0.67	0.67
US EE	130	18	7.22	7.21	0.37	0.21	0.67	-1.04	0.68	-1	0.92
US FF	96	18	5.33	5.25	-1.11	0.22	1.33	0.99	1.25	0.78	0.6
US GG	111	18	6.17	6.15	-0.43	0.21	0.98	0.04	0.99	0.06	0.84
UK 01	96	18	5.33	5.25	-1.11	0.22	0.91	-0.17	0.85	-0.34	0.86
UK 02	110	18	6.11	6.09	-0.47	0.21	0.57	-1.49	0.57	-1.48	0.86
UK 03	134	18	7.44	7.43	0.54	0.21	0.69	-0.93	0.68	-0.98	0.8
UK 04	146	18	8.11	8.14	1.05	0.21	0.58	-1.43	0.59	-1.41	0.8
UK 05	98	18	5.44	5.37	-1.01	0.22	1.3	0.93	1.2	0.66	0.73
UK 06	122	18	6.78	6.77	0.03	0.2	0.97	0.01	0.98	0.03	0.86
UK 07	150	18	8.33	8.39	1.23	0.21	0.69	-1	0.7	-0.96	0.79

With respect to Conversation items, the panellist judging the material to be at the highest CEFR level was UK07 (Table 11) with a fair average rating (i.e. the average adjusted according to the Rasch model) of 8.33 (B1+) while the lowest ratings were made by USFF at 5.25 (A2). Again, the ratings were acceptably consistent with just fit statistics for just one rater identified as potentially distorting (USCC at 2.04).

Table 12. Item Measurement Report for Conversation items (arranged by estimated item difficulty)

Item	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
Conversation 802 37	122	14	8.71	8.75	1.28	0.24	1.34	0.94	1.32	0.92

Conversation 802 255	98	14	7	7	-0.01	0.23	1.89	2.02	1.88	2
Conversation 802 404	121	14	8.64	8.68	1.23	0.24	1.41	1.1	1.43	1.14
Conversation 802 430	117	14	8.36	8.38	1	0.24	1.15	0.51	1.15	0.52
Conversation 802 540	112	14	8	8	0.73	0.23	0.96	0.01	0.97	0.04
Conversation 802 544	107	14	7.64	7.63	0.46	0.23	0.32	-2.48	0.31	-2.51
Conversation 802 545	119	14	8.5	8.53	1.12	0.24	1.73	1.73	1.69	1.67
Conversation 802 546	87	14	6.21	6.21	-0.6	0.23	0.39	-1.98	0.39	-2.01
Conversation 802 552	78	14	5.57	5.55	-1.09	0.23	0.62	-1.11	0.62	-1.11
Conversation 802 610	60	14	4.29	4.18	-2.17	0.26	0.77	-0.53	0.7	-0.64
Conversation 802 637	84	14	6	5.99	-0.76	0.23	1.13	0.46	1.13	0.45
Conversation 802 644	87	14	6.21	6.21	-0.6	0.23	1.63	1.53	1.64	1.54
Conversation 802 647	103	14	7.36	7.34	0.25	0.23	0.42	-1.93	0.42	-1.94
Conversation 802 744	63	14	4.5	4.43	-1.97	0.26	0.84	-0.33	0.77	-0.52
Conversation 802 750	65	14	4.64	4.59	-1.84	0.25	0.7	-0.78	0.66	-0.91
Conversation 802 755	91	14	6.5	6.5	-0.39	0.23	0.74	-0.61	0.74	-0.61
Conversation 802 756	77	14	5.5	5.48	-1.15	0.24	1.19	0.62	1.25	0.76
Conversation 802 767	109	14	7.79	7.78	0.57	0.23	0.42	-1.93	0.41	-1.97

The Conversation items (Table 12) ranged in difficulty from Conversation 802 610 with fair average rating of 4.18 (A1+) to Conversation 802 37 (8.75 or a high B1+). All items fell within the acceptable range for fit with values ranging from 0.32 (Conversation 802 544) to 1.89 (Conversation 802 255).

Table 13. Sentence Completion Item Ratings: All Facet Vertical Summary

```
+--------------------------------+
|Measr|+Raters        |+Item|CEFR |
|-----+--------------+-----+-----|
|   2 +              +     +(C1) |
|     |              |     |     |
|     |              |     | B2+ |
|     |              |     |     |
|     |              |     |     |
|     |              |     |     |
|     |              |     | --- |
|     |              |     |     |
|     |              |     |     |
|     |              | *   |  B2 |
|   1 +              +     +     |
|     |              | *** |     |
|     | UK 07  US AA |     |     |
|     |              | *   | --- |
|     | UK 04        | **  |     |
|     |              |     |     |
|     | UK 06        |     | B1+ |
|     | US EE        |     |     |
|     |              |     |     |
|     | UK 03        | *** |     |
*   0 * UK 05        *     * --- *
|     | US DD        | *** |     |
|     | US BB        |     |     |
|     |              |     |  B1 |
|     | UK 02  US GG |     |     |
|     |              | *   |     |
|     | UK 01  US CC | *   | --- |
|     | US FF        |     |     |
|     |              | **  |     |
|     |              | *   | A2+ |
|  -1 +              +     +     |
|     |              | *   |     |
|     |               |     | --- |
|     |               | **  |     |
|     |               | *   |     |
|     |               | **  |     |
|     |               |     |  A2 |
|     |               |     |     |
|     |               | *   |     |
|     |               | *   |     |
|  -2 +               +     +     |
|     |               | *   |     |
|     |               |     |     |
|     |               |     |     |
|     |               |     |     |
|     |               |     |     |
|     |               |     | --- |
|     |               | **  |     |
|     |               |     |     |
|     |               | *   |     |
|  -3 +               +     + (A1)|
|-----+---------------+-----+-----|
|Measr|+Raters        | * = |CEFR |
+---------------------------------+
```

Sentence Completion items (Table 13) ranged in difficulty from A1 to B2. Again there was a good spread of estimated difficulty with a number of items representing each CEFR level. There was a relatively narrow difference of just 1.5 logits or one CEFR level between the harshest and most lenient panellist.

Table 14. Rater Measurement Report for Sentence Completion items (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
US AA	229	30	7.63	7.69	0.81	0.15	1.66	2.29	1.73	2.42	0.48
US BB	183	30	6.1	5.87	-0.22	0.17	0.97	-0.03	0.87	-0.38	0.78
US CC	165	30	5.5	5.27	-0.57	0.17	0.84	-0.5	0.77	-0.72	0.87
US DD	190	30	6.33	6.02	-0.14	0.17	0.79	-0.75	0.74	-0.9	0.75
US EE	206	30	6.87	6.79	0.28	0.16	0.62	-1.63	0.65	-1.39	0.8
US FF	166	30	5.53	5.14	-0.66	0.18	1.43	1.4	1.29	0.96	0.68
US GG	175	30	5.83	5.56	-0.39	0.17	0.74	-0.93	0.79	-0.64	0.75
UK 01	162	30	5.4	5.18	-0.63	0.18	1.67	2.04	1.34	1.11	0.85
UK 02	176	30	5.87	5.56	-0.39	0.17	0.79	-0.72	0.81	-0.58	0.81
UK 03	196	30	6.53	6.41	0.07	0.16	0.88	-0.36	0.85	-0.47	0.82
UK 04	218	30	7.27	7.42	0.65	0.16	0.93	-0.19	0.92	-0.24	0.8
UK 05	190	30	6.33	6.27	0	0.16	1.31	1.14	1.25	0.9	0.83
UK 06	208	30	6.93	6.92	0.35	0.16	0.98	0.02	0.95	-0.09	0.77
UK 07	230	30	7.67	7.73	0.84	0.16	0.75	-1.01	0.77	-0.91	0.78

On Sentence Completion items (Table 14), the panellist judging the material to be at the highest CEFR level was again UK07 with a Fair Average rating of 7.33 (B1) while the lowest ratings were once again made by USFF at 5.14 (A2). Again, the ratings were acceptably consistent with fit statistics for just all raters falling within the acceptable range. The most misfitting panellist was UK01 with an Infit Mean Square of 1.67.

Table 15. Item Measurement Report for Sentence Completion items (arranged by estimated item difficulty)

Item	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
Sentence Completion 811 109	123	14	8.79	8.8	0.91	0.23	2.06	2.31	2.04	2.31
Sentence Completion 811 158	106	14	7.57	7.6	0.1	0.21	0.98	0.08	0.96	0
Sentence Completion 811 193	126	14	9	9.02	1.07	0.23	0.85	-0.3	0.86	-0.28
Sentence Completion 811 209	107	14	7.64	7.67	0.14	0.21	0.88	-0.21	0.84	-0.32
Sentence Completion 811 308	119	14	8.5	8.52	0.71	0.22	1.87	1.96	1.9	2.04
Sentence Completion 811 333	123	14	8.79	8.8	0.91	0.23	0.93	-0.08	0.91	-0.13
Sentence Completion 811 390	116	14	8.29	8.31	0.56	0.22	1.47	1.21	1.48	1.24
Sentence Completion 811 444	107	14	7.64	7.67	0.14	0.21	0.43	-1.87	0.43	-1.87
Sentence Completion 811 450	122	14	8.71	8.73	0.86	0.23	1.76	1.79	1.73	1.76
Sentence Completion 811 456	102	14	7.29	7.31	-0.08	0.21	1.46	1.25	1.52	1.36
Sentence Completion 811 721	64	14	4.57	4.49	-1.87	0.31	0.74	-0.46	0.65	-0.6
Sentence Completion 811 726	69	14	4.93	4.93	-1.47	0.26	0.73	-0.6	0.68	-0.62
Sentence Completion 811 729	83	14	5.93	5.93	-0.85	0.22	0.65	-1	0.77	-0.58
Sentence Completion 811 730	92	14	6.57	6.53	-0.52	0.21	0.73	-0.75	0.74	-0.72
Sentence Completion 811 732	73	14	5.21	5.02	-1.4	0.26	0.66	-0.82	0.81	-0.3
Sentence Completion 811 733	79	14	5.64	5.56	-1.05	0.23	0.87	-0.25	0.78	-0.48
Sentence Completion 811 737	66	14	4.71	4.58	-1.78	0.29	0.64	-0.75	0.62	-0.69
Sentence Completion 811 739	91	14	6.5	6.45	-0.57	0.21	0.57	-1.38	0.59	-1.31
Sentence Completion 811 740	74	14	5.29	5.2	-1.28	0.25	1.28	0.79	1.61	1.34
Sentence Completion 811 742	56	14	4	3.91	-2.65	0.44	0.55	-0.69	0.64	-0.59
Sentence Completion 811 745	69	14	4.93	4.93	-1.47	0.26	0.81	-0.36	0.82	-0.25
Sentence Completion 811 746	101	14	7.21	7.31	-0.08	0.21	1.52	1.38	1.48	1.29
Sentence Completion 811 750	74	14	5.29	5.11	-1.34	0.25	1.01	0.16	0.77	-0.42
Sentence Completion 811 753	101	14	7.21	7.24	-0.13	0.21	0.89	-0.22	0.9	-0.17
Sentence Completion 811 754	54	14	3.86	3.79	-2.86	0.47	0.54	-0.71	0.56	-0.81
Sentence Completion 811 756	83	14	5.93	5.75	-0.95	0.23	0.96	0.01	0.88	-0.21
Sentence Completion 811 757	54	14	3.86	3.91	-2.65	0.44	1.41	0.8	1.07	0.3
Sentence Completion 811 759	117	14	8.36	8.38	0.61	0.22	0.65	-0.95	0.66	-0.9
Sentence Completion 811 763	84	14	6	6.02	-0.8	0.22	0.67	-0.95	0.69	-0.88
Sentence Completion 811 764	59	14	4.21	4.31	-2.07	0.34	1.45	0.94	1.01	0.2

The Sentence Completion items (Table 15) had a greater range of estimated difficulty values than Conversations, Sentence Builds or Repeats, ranging from 3.79 (high A1: Sentence Completion 811 754) to 9.02 (B2: Sentence Completion 811 193). One item fell just outside the acceptable range for fit values: Sentence Completion 811 109 with an Infit Mean Square of 2.06.

Table 16. Dictation Item Ratings: All Facet Vertical Summary

```
+---------------------------------+
|Measr|+Raters        |+Item|CEFR |
|-----+---------------+-----+-----|
|   4 +               +     +(C1+)|
|     |               |     |     |
|     |               |     |  C1 |
|     |               |     |     |
|     |               |     |     |
|     |               |     |     |
|   3 +               + *   + --- |
|     |               |     |     |
|     |              |     |     |
|     |              |     | B2+ |
|     |              |     |     |
|     |              |     |     |
|   2 +              +     + --- |
|     |              |     |     |
|     | UK 04        | *   |     |
|     |              |     |  B2 |
|     | UK 06        | *   |     |
|     |              | *   |     |
|   1 + UK 07        + *   +     |
|     |              | *   | --- |
|     |              |     |     |
|     | US AA        | *   |     |
|     | US DD        |     | B1+ |
|     | UK 02        |     |     |
*   0 * US CC  US GG * *   *     *
|     | US BB        | **  | --- |
|     | US EE        |     |     |
|     | UK 03  UK 05 | *   |     |
|     |              | *   |  B1 |
|     |              |     |     |
|  -1 +              +     +     |
|     |              |     | --- |
|     |              | *   |     |
|     | UK 01        | *   | A2+ |
|     | US FF        |     |     |
|     |              | **  |     |
|  -2 +              + *   + --- |
|     |              | *   |     |
|     |              | *   |     |
|     |              | *   |     |
|     |              | *   |  A2 |
|     |              |     |     |
|  -3 +              +     +     |
|     |              | **  |     |
|     |              |     | --- |
|     |              |     |     |
|     |              |     |     |
|     |              |     | A1+ |
|  -4 +              + *   + (A1)|
|-----+--------------+-----+-----|
|Measr|+Raters       | * = |CEFR |
+--------------------------------+
```

Dictation items (Table 16) ranged in difficulty from A1 to B2+. However, there was only item at each of these two extreme levels and most items were between A1+ and B2. Panellists' judgements ranged over 3.41 logits (or approximately one and a half CEFR levels).

Table 17. Rater Measurement Report for Dictation items (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
US AA	174	24	7.25	7.33	0.5	0.21	0.76	-0.8	0.9	-0.25	0.86
US BB	158	24	6.58	6.61	-0.21	0.21	1.73	2.11	1.61	1.76	0.71
US CC	162	24	6.75	6.8	-0.03	0.21	0.81	-0.56	0.86	-0.36	0.86
US DD	172	24	7.17	7.24	0.41	0.21	0.8	-0.64	0.77	-0.7	0.86
US EE	154	24	6.42	6.41	-0.39	0.21	0.42	-2.44	0.42	-2.38	0.92
US FF	124	24	5.17	5.05	-1.74	0.22	0.9	-0.26	0.84	-0.29	0.83
US GG	161	24	6.71	6.75	-0.07	0.21	0.38	-2.66	0.37	-2.62	0.93
UK 01	128	24	5.33	5.2	-1.55	0.22	0.99	0.05	0.92	-0.1	0.89
UK 02	166	24	6.92	6.98	0.15	0.21	0.91	-0.22	0.93	-0.13	0.9
UK 03	152	24	6.33	6.3	-0.47	0.21	1.03	0.19	1	0.12	0.89
UK 04	202	24	8.42	8.46	1.67	0.21	1.8	2.35	1.82	2.32	0.74

UK 05	152	24	6.33	6.25	-0.52	0.21	1.61	1.87	1.63	1.85	0.93
UK 06	194	24	8.08	8.15	1.32	0.21	0.66	-1.3	0.65	-1.23	0.89
UK 07	84	24	7.67	7.76	0.93	0.21	0.92	-0.18	0.91	-0.22	0.85

As was the case for Conversation and Sentence Completion items, USFF proved to be the harshest rater for Dictation items (fair average 5.05 or A2) with UK04 emerging as the most lenient (fair average 8.46 or B1+). UK04 was also the most misfitting rater with an Infit Mean Square of 1.80 (Table 17).

Table 18. Item Measurement Report for Dictation items (arranged by estimated item difficulty)

Item	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
Dictation 812 155	149	14	10.64	10.67	3.03	0.3	2.25	2.68	2.07	2.11
Dictation 812 170	125	14	8.93	8.9	1.37	0.28	1.48	1.15	1.49	1.13
Dictation 812 312	113	14	8.07	8.13	0.45	0.27	1.34	0.97	1.27	0.78
Dictation 812 325	119	14	8.5	8.53	0.9	0.28	0.78	-0.48	0.8	-0.41
Dictation 812 501	121	14	8.64	8.66	1.06	0.28	1.39	1.02	1.46	1.13
Dictation 812 629	122	14	8.71	8.72	1.14	0.28	0.69	-0.75	0.68	-0.73
Dictation 812 979	98	14	7	7.03	-0.65	0.27	0.37	-1.99	0.35	-2.05
Dictation 812 982	99	14	7.07	7.1	-0.58	0.27	0.31	-2.32	0.32	-2.21
Dictation 812 983	77	14	5.5	5.43	-2.13	0.27	0.56	-1.24	0.54	-1.23
Dictation 812 984	104	14	7.43	7.46	-0.21	0.27	1.14	0.46	1.19	0.57
Dictation 812 985	79	14	5.64	5.57	-1.98	0.27	0.76	-0.58	0.81	-0.39
Dictation 812 986	65	14	4.64	4.65	-3.08	0.29	0.78	-0.52	0.76	-0.48
Dictation 812 988	106	14	7.57	7.61	-0.06	0.27	1.67	1.58	1.72	1.63
Dictation 812 992	128	14	9.14	9.09	1.61	0.28	0.54	-1.24	0.55	-1.13
Dictation 812 1212	70	14	5	4.98	-2.67	0.28	0.68	-0.75	0.74	-0.54
Dictation 812 1215	74	14	5.29	5.24	-2.36	0.28	1.35	0.91	1.38	0.95
Dictation 812 1228	72	14	5.14	5.11	-2.51	0.28	0.87	-0.2	0.8	-0.36
Dictation 812 1237	87	14	6.21	6.19	-1.42	0.26	0.57	-1.38	0.55	-1.42
Dictation 812 1240	89	14	6.36	6.35	-1.28	0.26	0.61	-1.19	0.59	-1.24
Dictation 812 1296	55	14	3.93	3.8	-3.95	0.31	0.64	-1.01	0.55	-0.62
Dictation 812 1303	63	14	4.5	4.5	-3.25	0.29	1.17	0.58	1.04	0.23
Dictation 812 1307	81	14	5.79	5.72	-1.83	0.27	0.55	-1.37	0.54	-1.38
Dictation 812 1321	105	14	7.5	7.54	-0.13	0.27	0.57	-1.19	0.57	-1.17
Dictation 812 1324	82	14	5.86	5.8	-1.76	0.27	2.58	3.26	2.62	3.25

These Dictation items (Table 18) were judged to range in difficulty from 3.80 (Dictation 812 1296: a high A1) to 10.67 (B2+: Dictation 812 155). The latter was one of two items – the other being Dictation 812 1324 – with Infit Mean Square values greater than 2.0. This is indicative of a lack of consistency on the part of panellists in rating these items with a potentially distorting effect on the resulting measurement scale.

Overall, these objectively scored items showed that panellists were consistent in their judgements. There was a broad consensus on the relative difficulty of the items and this generally accorded with the observed order of difficulty in the operational test. On the other hand, there was quite a wide range when it came to judging the CEFR level of the items. Some panellists such as USFF and UK01 were consistently harsh in their judgements while others such as UK04 and UK07 were consistently lenient across item types.

Correlations between panellists' judgements and observed item difficulties

Having obtained difficulty estimates for each item, based on the panel's judgements, these estimates could be compared with the observed difficulty of the items obtained through routine administration of the test. One would anticipate a positive correlation between the panellists estimates of item difficulty and the observed item difficulty. This proved to be the case: the correlation between the Rasch difficulty estimates based on the panellists judgements and difficulty estimates based on test performance for these items are shown in Table 19.

Table 19. Correlations between panel judgements and observed item difficulty for dichotomous task types

Task Type	Correlation
Repeats & Sentence Builds	0.776
Conversations	0.909
Sentence Completion	0.888
Dictation	0.773

The correlations between the panellists' judgements and the observed difficulty of the Conversation items (at 0.909) and Sentence Completion items (0.888) were relatively high (Table 19), showing that the panellists' judgements were very consistent with the observed difficulty of the items for these task types. The figures for Repeats & Sentence Builds and Dictation items were lower (0.776 and 0.773 respectively), but nonetheless indicative of a moderately strong relationship.

Table 20. CEFR: VEPT links based on panellists judgements of item difficulty

	Sentence Builds and Repeats		**Conversations**		**Sentence Completion**		**Dictation**	
CEFR	Thurstone Thresholds (Logits)	VEPT (20-70)	Thurstone Thresholds (Logits)	VEPT (20-70)	Thurstone Thresholds (Logits)	VEPT (20-70)	Thurstone Thresholds (Logits)	VEPT (20-70)
A1+	-2.82	11.78	-2.31	-0.86			-3.80	38.06
A2	-2.79	12.09	-2.21	1.08	-2.43	23.22	-3.69	38.40
A2+	-1.32	27.27	-0.82	28.00	-0.87	39.05	-1.65	44.57
B1	-1.28	27.68	-0.68	30.71	-0.82	39.55	-1.47	45.12
B1+	0.06	41.52	0.64	56.28	0.23	50.20	0.10	49.87
B2	0.10	41.93	0.81	59.57	0.45	52.44	0.51	51.11
B2+	1.54	56.80	2.11	84.75	1.69	65.01	2.48	57.07
C1	1.68	58.25	2.40	90.37	1.84	66.54	2.60	57.43

The results of the panellists' judgements provide a somewhat inconsistent picture of the appropriate CEFR-VEPT relationship. This is partly because, although the panellists found a similar spread of difficulty among items of different types, the range of difficulty differed considerably on the VEPT scale according to item type. All Dictation items considered by the panellists fell within a range of 25 points on the VEPT scale while the Conversation items covered a range of 79 points. Note that although scores are *reported* on a scale ranging from 20 to 70, they may in fact cover a wider range than this.

A related issue is the discrepancy between the machine scoring used in the test (which locates each response on a scale) and the treatment of items as dichotomous for the purpose of the panel. Test takers giving what the panellists would regard as incorrect responses would not, in fact score 0 on the item and those giving 'correct' responses would not score 1 on the operational test. The compromise of treating items as dichotomous was unavoidable for the reasons given above, but limit the comparability of panellists' judgements and observed scores.

Taking an average across task types and rounding up to the nearest integer suggests cut scores of 21 for A2, 36 for B1, 52 for B2 and 69 for C1.

Performance Rating: Read Aloud, Passage Reconstruction, Summary and Opinion

Table 21. VEPT Item Ratings: All Facet Vertical Summary for performance items

```
+-----------------------------------------------+
 |Measr|-Perf.   |-Raters                 |CEFR |
 |-----+---------+------------------------+-----|
 |    6 +         +                        +(C2+)|
 |     | **      |                        |     |
 |     | *       |                        |     |
 |     | *       |                        |     |
 |    5 + **      +                        +     |
 |     | *       |                        | C2  |
 |     | *       |                        |     |
 |     | ***     |                        |     |
 |    4 + ***     +                        +     |
 |     |         |                        |     |
 |     | *       |                        | --- |
 |     |         |                        | C1+ |
 |    3 + ***     +                        +     |
 |     | *       |                        | --- |
 |     | ***     |                        | C1  |
 |     | ****    |                        | B2+ |
 |    2 + ***     +                        +     |
 |     | *       |                        | --- |
 |     | ****    |                        | B2  |
 |     | ****    |                        |     |
 |    1 + ***     +                        + --- |
 |     |         | USAA                   | B1+ |
 |     | ****** | USDD  USFF  USGG       |     |
 |     | **      | USCC  USEE             | --- |
 *    0 * ****** * USBB                   *     *
 |     | **      | UK05  UK06  UK07       | B1  |
 |     | **      | UK01  UK02  UK03  UK04 | --- |
 |     | *****   |                        |     |
 |   -1 + ****    +                        + A2+ |
 |     | ****    |                        |     |
 |     | ****    |                        | --- |
 |     | ****** |                        |     |
 |   -2 + ***     +                        + A2  |
 |     | ***     |                        |     |
 |     | *****   |                        |     |
 |     | ****** |                        | --- |
 |   -3 + **      +                        + A1+ |
 |     | ***     |                        |     |
 |     | **      |                        | --- |
 |     | *       |                        |     |
 |   -4 +         +                        +     |
 |     | *       |                        | A1  |
 |     |        |                        |     |
 |     | *      |                        | --- |
 |  -5 +        +                        +(A0) |
 |-----+--------+------------------------+-----|
 |Measr| * = 1  |-Raters                 |CEFR |
 |-----+--------+------------------------+-----|
```

109 performances on individual test items were rated by panellists (Table 21). The sample performances were divided into two sets: one for each panel. These two data sets overlapped so that 55 of the performances were rated by panellists in both groups. 14 were rated only by UK panellists and 40 only by US panellists. The second column in Table 21 (headed 'Perf.') shows the distribution of test taker ability estimates. It can be seen that there was a rather greater range in levels for these performances than for the test material, with ratings ranging from A1 to C2. Again, this reflects the wide range of proficiency measured by the VEPT.

Table 22. Rater Measurement Report for performance items (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
UK02	540	69	7.83	7.66	-0.58	0.11	0.34	-4.85	0.36	-4.43	0.98
UK04	536	69	7.77	7.61	-0.53	0.11	0.33	-5.05	0.34	-4.62	0.97
UK01	525	69	7.61	7.45	-0.40	0.11	0.96	-0.15	0.99	-0.01	0.97
UK03	524	69	7.59	7.44	-0.39	0.11	0.78	-1.22	0.81	-1.04	0.97
UK06	519	69	7.52	7.37	-0.32	0.11	0.57	-2.76	0.58	-2.59	0.97
UK07	510	69	7.39	7.25	-0.22	0.11	0.53	-3.09	0.56	-2.74	0.97
UK05	505	69	7.32	7.18	-0.15	0.11	0.57	-2.79	0.54	-2.93	0.97
USBB	541	75	7.21	6.89	0.12	0.12	1.64	2.91	1.71	3.14	0.87
USCC	572	78	7.33	6.86	0.14	0.10	1.28	1.65	1.41	2.29	0.88
USEE	569	78	7.29	6.83	0.17	0.10	1.16	1.01	1.18	1.10	0.88
USGG	546	78	7.00	6.58	0.40	0.10	1.30	1.76	1.57	3.03	0.87
USDD	537	78	6.88	6.48	0.50	0.10	1.05	0.33	1.08	0.51	0.90
USFF	531	78	6.81	6.41	0.56	0.10	0.97	-0.13	1.04	0.28	0.92
USAA	517	78	6.63	6.25	0.71	0.10	1.15	0.90	1.19	1.09	0.88

In Table 22, the ‘T.Count’ column shows the number of ratings made by each panellist. The column labelled ‘Observed Average’ shows the average of the ratings awarded to the test taker performances by each rater. The ‘Fair Average’ column shows that the harshest rater, USAA, would, it is estimated, tend to judge an average level performance (i.e. 7.1 or B1) as 6.25 (A2+) while most lenient rater, UK02 would estimate it at 7.66 (a high B1). In this case, there appears to be a small, but clear divergence between the UK and US panels with the UK panellists on average consistently awarding higher ratings by 0.81 points on the CEFR scale. The most severe UK panellist (UK05) awarded higher scores than the least severe of the US judges (USCC).

Table 22 suggests that panellists in both groups were consistent in their judgements. USBB was the least consistent, the Infit Mean Square value of 1.65 for this rater falling outside the 0.5 – 1.5 range that is said to be productive for measurement. The UK panel were in closer agreement with both UK02 and UK04 being more consistent than the model would predict and falling below the 0.5 criterion for productive measurement. Although low Infit Mean Square values may suggest conservatism in the use of the scales, in fact the American panel generally made use of a narrower range of categories than the UK group (although comparisons are complicated by the fact that the two panels did not rate precisely the same sets of performances). A high proportion of scores for both panels fell into the A2 to B1+ range, but 48% of scores for the US panel were in this range compared with 34% for the UK panel which produced proportionately more scores both below and above this level.

The SR- RoR correlations (PtBis) show even greater consistency among raters than was observed for the dichotomously scored items.

Table 23. Item Measurement Report for performance items (arranged by estimated person ability)

Items	*T.Score*	*T.Count*	*Obs.Avge*	*Fair.Avge*	*Measure*	*S.E.*	*InfitMS*	*InfitZ*	*OutfitMS*	*OutfitZ*
READ ALOUD 133723907	21	14	1.50	1.44	-5.83	0.32	1.78	1.92	2.39	2.10
PASSAGE REC. 113938120	22	14	1.57	1.51	-5.74	0.31	1.12	0.47	1.21	0.58
READ ALOUD 145229122	25	14	1.79	1.74	-5.48	0.28	1.54	1.75	1.63	1.65
READ ALOUD 150615608	27	14	1.93	1.90	-5.32	0.28	1.94	2.76	2.34	3.29
OPINION 114303580	2	1	2.00	2.14	-5.10	0.96	1.00	0.00	1.00	0.00
SUMMARY 113938120	2	1	2.00	2.14	-5.10	0.96	1.00	0.00	1.00	0.00
READ ALOUD 144824512	34	14	2.43	2.46	-4.81	0.27	2.01	2.48	2.04	2.40
SUMMARY 114086939	21	7	3.00	2.68	-4.59	0.41	1.21	0.51	1.21	0.51
READ ALOUD 138066569	40	14	2.86	2.89	-4.36	0.28	0.28	-2.39	0.25	-2.46
PASSAGE REC. 113980575	41	14	2.93	2.95	-4.28	0.28	1.48	1.14	1.54	1.21
PASSAGE REC. 113545578	23	7	3.29	2.99	-4.24	0.41	0.16	-1.81	0.15	-1.82
OPINION 113980575	3	1	3.00	3.10	-4.10	1.10	1.00	0.00	1.00	0.00
SUMMARY 113567488	24	7	3.43	3.12	-4.07	0.41	0.47	-0.86	0.46	-0.90
PASSAGE REC. 114303580	46	14	3.29	3.28	-3.88	0.28	1.25	0.69	1.36	0.90
SUMMARY 114086939	20	6	3.33	3.70	-3.40	0.44	0.82	-0.03	0.81	-0.05
READ ALOUD 145646818	57	14	4.07	4.07	-3.05	0.27	1.28	0.86	1.28	0.86
READ ALOUD 150444971	57	14	4.07	4.07	-3.05	0.27	1.51	1.40	1.48	1.32
OPINION 113566437	4	1	4.00	4.13	-2.99	0.98	1.00	0.00	1.00	0.00
PASSAGE REC. 113231167	28	7	4.00	4.39	-2.74	0.37	2.03	1.98	2.02	1.97
READ ALOUD 150344641	63	14	4.50	4.52	-2.62	0.27	2.07	2.38	2.05	2.30
READ ALOUD 134390647	64	14	4.57	4.59	-2.54	0.27	1.95	2.13	1.93	2.07
READ ALOUD 147462809	65	14	4.64	4.66	-2.47	0.27	1.23	0.69	1.23	0.69
PASSAGE REC. 113566437	31	7	4.43	4.80	-2.33	0.37	0.85	-0.18	0.85	-0.17
READ ALOUD 149424739	67	14	4.79	4.80	-2.32	0.27	1.40	1.04	1.39	1.01
SUMMARY 112933154	36	7	5.14	4.80	-2.32	0.39	0.11	-2.29	0.11	-2.30
OPINION 113133431	64	13	4.92	4.93	-2.19	0.28	0.70	-0.68	0.69	-0.71
OPINION 113249440	5	1	5.00	5.11	-1.99	1.05	1.00	0.00	1.00	0.00
READ ALOUD 139755741	72	14	5.14	5.14	-1.95	0.27	1.02	0.17	1.02	0.19
READ ALOUD 149653238	72	14	5.14	5.14	-1.95	0.27	1.81	1.76	1.81	1.74
OPINION 114086055	29	6	4.83	5.22	-1.87	0.42	2.25	1.68	2.31	1.73
PASSAGE REC. 113147938	77	14	5.50	5.49	-1.59	0.27	1.38	0.99	1.42	1.06
PASSAGE REC. 113325003	41	7	5.86	5.48	-1.59	0.37	0.39	-1.47	0.39	-1.47
PASSAGE REC. 113937729	42	7	6.00	5.62	-1.45	0.37	0.34	-1.72	0.34	-1.74
READ ALOUD 113129020	79	14	5.64	5.63	-1.44	0.27	0.53	-1.35	0.53	-1.33
SUMMARY 113545832	32	6	5.33	5.73	-1.34	0.42	1.78	1.20	1.80	1.21
OPINION 113987501	75	13	5.77	5.74	-1.33	0.28	1.18	0.57	1.22	0.65
PASSAGE REC. 113249440	81	14	5.79	5.77	-1.30	0.26	0.35	-2.21	0.36	-2.14
READ ALOUD 136792942	82	14	5.86	5.84	-1.23	0.26	0.78	-0.52	0.83	-0.35
SUMMARY 113231167	6	1	6.00	6.13	-0.96	0.97	1.00	0.00	1.00	0.00
SUMMARY 113681644	6	1	6.00	6.13	-0.96	0.97	1.00	0.00	1.00	0.00
PASSAGE REC. 112932983	46	7	6.57	6.17	-0.93	0.36	0.35	-1.50	0.34	-1.51
OPINION 113137749	85	13	6.54	6.53	-0.59	0.27	0.63	-0.97	0.63	-0.96
READ ALOUD 144892612	92	14	6.57	6.57	-0.55	0.26	0.78	-0.50	0.77	-0.52
OPINION 113134006	49	7	7.00	6.60	-0.53	0.36	0.02	-3.52	0.02	-3.52
SUMMARY 114208530	37	6	6.17	6.61	-0.52	0.39	0.97	0.12	0.97	0.11

SUMMARY 114418055	37	6	6.17	6.61	-0.52	0.39	0.97	0.12	0.97	0.11
OPINION 113939000	50	7	7.14	6.74	-0.40	0.36	0.40	-1.17	0.40	-1.16
SUMMARY 112932083	38	6	6.33	6.78	-0.36	0.39	1.87	1.46	1.87	1.47
PASSAGE REC. 113148218	45	7	6.43	6.83	-0.32	0.36	0.92	0.02	0.93	0.02
PASSAGE REC. 113681644	99	14	7.07	7.06	-0.10	0.25	0.56	-1.22	0.57	-1.19
SUMMARY 113147938	7	1	7.00	7.13	-0.04	0.96	1.00	0.00	1.00	0.00
SUMMARY 114565356	7	1	7.00	7.13	-0.04	0.96	1.00	0.00	1.00	0.00
SUMMARY 113483623	94	13	7.23	7.21	0.04	0.26	0.36	-2.04	0.36	-2.02
PASSAGE REC. 113137482	48	7	6.86	7.26	0.08	0.36	1.04	0.25	1.03	0.25
READ ALOUD 145489254	102	14	7.29	7.27	0.09	0.25	2.65	3.12	2.72	3.20
OPINION 113848260	96	13	7.38	7.36	0.17	0.26	0.70	-0.70	0.69	-0.75
READ ALOUD 133454852	104	14	7.43	7.41	0.22	0.25	0.79	-0.46	0.82	-0.36
OPINION 113485613	57	7	8.14	7.69	0.46	0.34	0.34	-1.57	0.34	-1.56
OPINION 113901154	44	6	7.33	7.81	0.56	0.39	0.60	-0.57	0.60	-0.56
READ ALOUD 145446276	113	14	8.07	8.06	0.76	0.24	0.68	-0.81	0.67	-0.85
READ ALOUD 113459555	114	14	8.14	8.13	0.82	0.24	0.53	-1.33	0.53	-1.33
SUMMARY 113148218	8	1	8.00	8.14	0.83	0.90	1.00	0.00	1.00	0.00
SUMMARY 114000851	8	1	8.00	8.14	0.83	0.90	1.00	0.00	1.00	0.00
READ ALOUD 145015503	54	7	7.71	8.16	0.84	0.35	0.55	-0.86	0.55	-0.86
READ ALOUD 133455047	116	14	8.29	8.27	0.94	0.24	1.01	0.17	1.01	0.16
READ ALOUD 114559857	118	14	8.43	8.41	1.05	0.24	0.88	-0.18	0.90	-0.13
READ ALOUD 150537521	118	14	8.43	8.41	1.05	0.24	0.36	-2.01	0.37	-1.96
OPINION 113939092	110	13	8.46	8.44	1.06	0.24	0.49	-1.37	0.49	-1.36
PASSAGE REC. 114284986	121	14	8.64	8.62	1.21	0.23	2.22	2.35	2.30	2.51
PASSAGE REC. 113918700	65	7	9.29	8.77	1.32	0.31	0.24	-1.64	0.25	-1.61
READ ALOUD 148175406	123	14	8.79	8.76	1.32	0.23	0.79	-0.42	0.82	-0.34
PASSAGE REC. 114000851	124	14	8.86	8.83	1.37	0.23	0.38	-1.85	0.41	-1.75
OPINION 113544349	50	6	8.33	8.86	1.39	0.36	0.56	-0.70	0.56	-0.70
OPINION 113015962	66	7	9.43	8.89	1.42	0.30	0.14	-2.21	0.14	-2.22
READ ALOUD 143415231	127	14	9.07	9.03	1.52	0.22	0.68	-0.76	0.71	-0.66
OPINION 113137482	9	1	9.00	9.16	1.62	0.86	1.00	0.00	1.00	0.00
PASSAGE REC. 113395629	61	7	8.71	9.21	1.65	0.33	0.65	-0.49	0.65	-0.48
READ ALOUD 148291812	130	14	9.29	9.23	1.66	0.21	1.52	1.26	1.57	1.35
PASSAGE REC. 114565356	132	14	9.43	9.36	1.75	0.21	0.59	-1.10	0.63	-0.94
READ ALOUD 114559597	62	7	8.86	9.37	1.76	0.33	0.40	-1.12	0.40	-1.12
READ ALOUD 148526909	133	14	9.50	9.43	1.80	0.21	0.62	-1.04	0.70	-0.71
READ ALOUD 138725730	63	7	9.00	9.53	1.86	0.32	0.78	-0.20	0.75	-0.24
SUMMARY 113778254	127	13	9.77	9.67	1.94	0.21	0.25	-2.86	0.25	-2.71
PASSAGE REC. 112285729	64	7	9.14	9.70	1.96	0.32	0.41	-1.07	0.39	-1.11
READ ALOUD 148324653	137	14	9.79	9.70	1.97	0.20	1.00	0.11	1.08	0.34
OPINION 114284986	10	1	10.00	10.22	2.24	0.73	1.00	0.00	1.00	0.00
SUMMARY 113395629	10	1	10.00	10.22	2.24	0.73	1.00	0.00	1.00	0.00
PASSAGE REC. 112001388	68	7	9.71	10.39	2.33	0.29	0.53	-0.87	0.47	-1.00
READ ALOUD 114303539	150	14	10.71	10.69	2.46	0.19	0.31	-2.81	0.32	-2.70
PASSAGE REC. 112554704	70	7	10.00	10.73	2.48	0.28	0.79	-0.28	0.73	-0.41
READ ALOUD 114411111	71	7	10.14	10.89	2.56	0.27	1.18	0.51	1.09	0.35
PASSAGE REC. 116043195	153	14	10.93	10.93	2.58	0.19	1.03	0.21	1.02	0.17
SUMMARY 113015961	61	6	10.17	11.01	2.62	0.29	0.61	-0.69	0.62	-0.65
PASSAGE REC. 112929146	155	14	11.07	11.10	2.65	0.20	1.33	1.02	1.23	0.75
READ ALOUD 114562616	156	14	11.14	11.18	2.69	0.20	0.65	-1.11	0.63	-1.12
OPINION 116043195	11	1	11.00	11.25	2.72	0.68	1.00	0.00	1.00	0.00
SUMMARY 114398286	146	13	11.23	11.26	2.73	0.21	0.84	-0.37	0.77	-0.56
SUMMARY 111851382	84	7	12.00	11.34	2.77	0.31	0.29	-1.79	0.29	-1.71
SUMMARY 114823079	63	6	10.50	11.36	2.78	0.29	1.18	0.52	1.19	0.53
READ ALOUD 139313287	162	14	11.57	11.67	2.94	0.21	0.36	-2.39	0.38	-2.09
PASSAGE REC. 112001117	163	14	11.64	11.75	2.98	0.21	0.97	0.03	0.90	-0.15
OPINION 112937899	156	13	12.00	12.12	3.22	0.24	0.63	-0.94	0.53	-1.03
SUMMARY 112001388	12	1	12.00	12.16	3.25	0.81	1.00	0.00	1.00	0.00
OPINION 113315468	158	13	12.15	12.28	3.33	0.25	1.23	0.65	0.99	0.15
READ ALOUD 138266868	174	14	12.43	12.53	3.59	0.27	1.05	0.26	0.63	-0.54
READ ALOUD 146262892	174	14	12.43	12.53	3.59	0.27	0.99	0.15	0.60	-0.61

READ ALOUD 124355681	85	7	12.14	12.56	3.63	0.33	1.22	0.55	1.00	0.20
PASSAGE REC. 113317749	91	7	13.00	12.90	4.28	0.77	0.00	-1.74	0.00	-1.76

Table 23 shows how these performances are located on the measurement scale for these items. In this case, fifteen performance ratings had Infit Mean Square values above the level that is considered productive for measurement, six of these being unproductive (>1.5), and six potentially distorting (*Read Aloud 145489254, Opinion 114086055, Passage Reconstruction 114284986, Read Aloud 150344641, Passage Reconstruction 113231167, Read Aloud 144824512*) Again, the small numbers involved suggest that removing these cases is unnecessary and all difficulty estimates were retained in the subsequent analysis.

Correlations between panellists' judgements and test takers' machine-generated scores

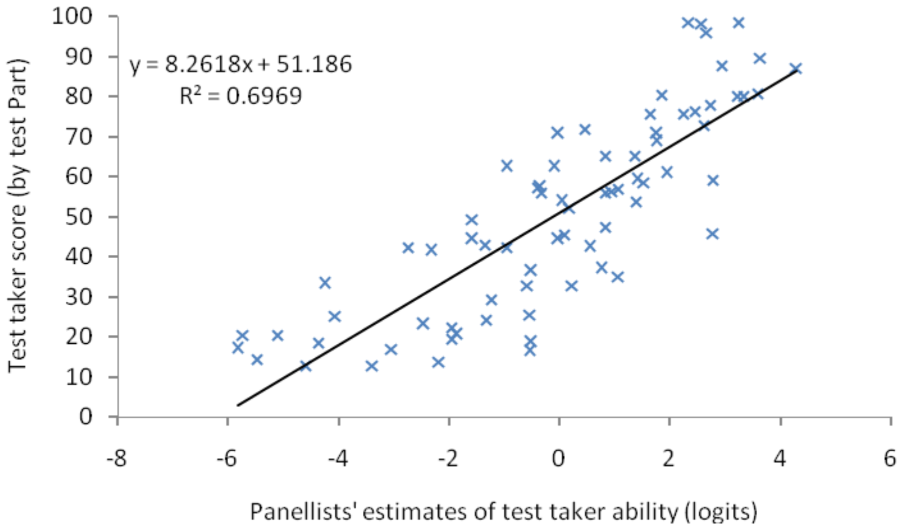

Figure 3. Scatterplot of panellists' judgements of test taker performance on performance items by test takers' machine-generated scores

Comparing the pooled judgement of the panellists of test taker performance on these items with the test takers' scores gives a correlation of .835, indicating substantial agreement between the two. This relationship is represented visually in Figure 3.

Table 24. CEFR: VEPT links based on panellists judgements of test takers' performances on performance items

CEFR	Rasch Thurstone Thresholds	VEPT (20-70)
A0+	-5.21	8.1
A1	-5.19	8.3
A1+	-3.24	24.4
A2	-2.92	27.1
A2+	-1.32	40.3
B1	-0.81	44.5
B1+	0.49	55.2
B2	0.96	59.1

B2+	2.02	67.9
C1	2.38	70.8
C1+	2.85	74.7
C2	3.11	76.9
C2+	6.81	107.4

The evidence from these performance items suggests that test takers will need to be at A1 level to register a score over 20, that a score of 28 or higher is consistent with the A2 level, 45 or higher for B1 level and 60 or more for B2. From this perspective a score of 70 would not quite be enough for the test taker to be placed at C1 level.

Body of work: Performance on the test as a whole

Finally, the panellists considered the performance of individual test takers on the test as a whole. They assigned a CEFR level to the test taker's responses to each of the eight test Parts. Thus each panellist made eight CEFR ratings for each test taker and these were used in arriving at the ability estimates given below.

Table 25. VEPT Item Ratings: All Facet Vertical Summary for Body of Work judgements

```
+---------------------------------------------------+
|Measr|-Raters                     |+Testtaker|CEFR |
|-----+----------------------------+----------+-----|
|   3 +                            +          +(C2) |
|     |                            |          | --- |
|     |                            |          |     |
|     |                            |          | C1  |
|     |                            |          |     |
|     |                            |          | --- |
|     |                            | VEPT=69  |     |
|   2 +                            +          + B2+ |
|     |                            |          |     |
|     |                            |          | --- |
|     |                            |          |     |
|     |                            |          |  B2 |
|     |                            | VEPT=60  |     |
|     |                            |          |     |
|   1 +                            +          + --- |
|     | US FF                      |          |     |
|     |                            |          |  B1+|
|     |                            |          |     |
|     |                            |          |     |
|     | UK 05  US GG               | VEPT=51  | --- |
|     | US AA  US BB  US DD  US EE |          |     |
*   0 * UK 06  US CC               *          *  B1 *
|     | UK 01  UK 03               |          |     |
|     |                            |          | --- |
|     | UK 04                      |          |     |
|     | UK 02  UK 07               |          |     |
|     |                            | VEPT=50  |  A2+|
|     |                            |          |     |
|  -1 +                            +          +     |
|     |                            |          | --- |
|     |                            |          |     |
|     |                            | VEPT=40  |  A2 |
|     |                            |          |     |
|     |                            | VEPT=33  |     |
|     |                            |          | --- |
|  -2 +                            +          +  A1+|
|     |                            |          |     |
|     |                            |          | --- |
|     |                          |          |     |
|     |                          |          |  A1 |
|     |                          |          | --- |
|     |                          | VEPT=29  |     |
|  -3 +                          + VEPT=22  +  A0+|
|     |                          |          |     |
|     |                          |          |     |
|     |                          |          | --- |
|     |                          |          |     |
|     |                          |          |     |
|     |                          |          |     |
|  -4 +                          +          + (A0)|
|-----+--------------------------+----------+-----|
|Measr|-Raters                   |+Testtaker|CEFR |
+-------------------------------------------------+
```

Table 25 displays the ability estimates for eight test takers. Each test taker's overall VEPT score is shown in the third column: VEPT=22 represents a VEPT test taker with an overall score on the test of 22 points. This individual is judged to be at A0+ while the test taker scoring 69 points (VEPT=69) is located at the borderline between B2+ and C1.

Table 26. Rater Measurement Report for Body of Work judgements (arranged by estimated rater harshness)

Raters	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtBis
UK 01	294	48	6.12	6.06	-0.08	0.11	1.47	2.20	1.52	2.34	0.93
UK 02	339	48	7.06	6.79	-0.62	0.11	1.64	2.92	1.51	2.35	0.92
UK 03	300	48	6.25	6.16	-0.15	0.11	1.27	1.36	1.35	1.68	0.96
UK 04	325	48	6.77	6.56	-0.45	0.11	1.42	2.10	1.46	2.17	0.89
UK 05	269	48	5.60	5.64	0.25	0.12	0.77	-1.11	0.76	-1.17	0.95
UK 06	283	48	5.90	5.88	0.06	0.11	1.17	0.85	1.14	0.71	0.92
UK 07	166	32	5.19	6.73	-0.57	0.13	1.07	0.36	0.94	-0.17	0.82
US AA	289	48	6.02	5.86	0.08	0.12	0.77	-1.09	0.78	-1.02	0.94
US BB	283	48	5.90	5.76	0.16	0.12	0.50	-2.80	0.51	-2.72	0.95
US CC	299	48	6.23	6.03	-0.06	0.12	0.57	-2.40	0.57	-2.35	0.94
US DD	283	48	5.90	5.76	0.16	0.12	0.53	-2.59	0.56	-2.38	0.94
US EE	284	48	5.92	5.77	0.14	0.12	0.63	-1.94	0.65	-1.80	0.95
US FF	231	48	4.81	4.80	0.86	0.12	1.39	1.76	2.32	3.84	0.87
US GG	277	48	5.77	5.66	0.24	0.12	0.60	-2.06	0.70	-1.49	0.93

As shown by the 'T.Count' column in Table 26, all but one of the panellists made 48 ratings. The exception was one of the UK panellists who was unavailable to rate the last two test takers. Both groups rated test takers VEPT=22, VEPT=40, VEPT=60 and VEPT=69. The UK panellists also rated VEPT=29 and VEPT=51, while the US panellists rated VEPT=33 and VEPT=50. The UK panellists again proved to be somewhat more lenient than the US group (Table 26), although on this occasion the harshest UK rater (UK05) was the second most severe overall. The Infit Mean Square results again indicate good levels of consistency: only UK02 (with an Infit Mean Square of1.6) falling outside the 'productive for measurement' range of 0.5 to 1.5.

Table 27. Person Measurement Report for Body of Work judgements (arranged by estimated person ability)

Test taker	T.Score	T.Count	Obs.Avge	Fair.Avge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ

VEPT=22	235	112	2.1	1.99	-3.02	0.08	1.43	2.98	1.76	4.21
VEPT=29	158	56	2.82	2.32	-2.83	0.09	1.76	4.01	1.74	3.88
VEPT=33	229	56	4.09	4.53	-1.78	0.1	0.48	-3.5	0.47	-3.58
VEPT=40	560	112	5	5.05	-1.45	0.08	0.84	-1.08	0.91	-0.56
VEPT=50	323	56	5.77	6.06	-0.68	0.12	0.59	-2.24	0.58	-2.28
VEPT=51	432	56	7.71	7.38	0.26	0.11	0.91	-0.45	0.91	-0.47
VEPT=60	912	104	8.77	8.83	1.24	0.08	1.06	0.45	1.05	0.39
VEPT=69	1073	104	10.32	10.38	2.2	0.07	0.72	-2.34	0.75	-2.12

Table 27 also shows that the panellists were able to arrive at consistent decisions on the CEFR level of these test takers. Only the results for VEPT=29 fall outside the acceptable range for productive measurement.

Table 28. CEFR: VEPT links based on panellists judgements of test takers' performances on the test as a whole

CEFR level	Rasch Thurstone Thresholds	VEPT (20-70)
A0	low	-
A0+	-2.75	27.45
A1	-2.73	27.62
A1+	-2.08	33.10
A2	-1.87	34.87
A2+	-0.98	42.37
B1	-0.31	48.02
B1+	0.32	53.33
B2	0.86	57.89
B2+	1.69	64.88
C1	2.12	68.51

Table 28 shows the range of the logit scale associated with each CEFR level, the Rasch Thurstone Thresholds indicating the boundaries between one level and the next and its corresponding VEPT score. Any rating below -2.75 on the logit measure is categorised as A0, from -2.75 to -2.73 is categorised as A0+, from -2.73 to -2.08 is A1 and so on. It can be seen that the test takers were ranked in the same order by both the VEPT and the panellists. The correlation between VEPT scores and Rasch measures for these test takers was .984. Using the current CEFR-VEPT recommendations, all but one of the eight test takers was placed into the same CEFR by the consensus of panellists and by VEPT score. The exception was VEPT=33, classified as A2 by the VEPT, but as A1 by the panellists.

Conclusions

Cizek and Bunch acknowledge that 'different methods [of standard setting] will yield different results, and there is no way to determine that one method or another produced the wrong results' (p.63). In the absence of a definitive methodology, it is appropriate to employ multiple methods and aggregate the results.

Of the approaches adopted here, the Body of Work is the most comprehensive as it takes account of performance on the test as a whole. On the other hand, the number of Body of Work judgements that can be made in a short time is necessarily limited and may not be representative. The task centred variation on the Basket method and person centred sample rating method both yielded more data and, because they involved consideration of individual test items, could provide a basis for embedding CEFR judgements in routine test production.

Inevitably, the three approaches each suggested a somewhat different relationship between the VEPT and the CEFR. The Basket method and performance sample method - both based on individual test items - suggested that any test taker registering a score over 20 would be at least at the A1 level. On the other hand, the two lowest scoring test takers considered in the Body of Work activity were both judged to be at the A0 level (although their scores were 22 and 29). The Basket method suggested that a score of 58 or higher might equate to C1 on the CEFR while the performance sample evidence suggested that the cut point for C1 was above the maximum VEPT score of 70.

Taking a simple average of the results obtained across the three methods and rounding up to the nearest integer (or whole score point) gives the following recommendations on VEPT-CEFR correspondences (Table 29).

Table 29. Recommended Mapping of CEFR Levels with Versant English Placement Test Overall Scores

VEPT (20-70)	CEFR (A1-C1)
20-23	<A1
24-33	A1
34-45	A2
46-56	B1
57-65	B2
66-70	C1

Table 29 represents the panellists' recommendations on the vertical relationship between the VEPT and the CEFR levels.

This table needs to be understood in context. First, the horizontal relationship between test and framework must also be taken into account. The VEPT does not involve all of the features of language covered in the CEFR and the overall picture that the test scores provide of a learner's CEFR level may mask strengths or weaknesses in specific areas. The specification exercise suggested that sociolinguistic competences are not addressed by the VEPT and functional uses of language that come to the fore at the higher levels of the CEFR are not reflected in the test material. The test provides, for example, no opportunities to manage an interaction or organise an extended report. Panellists made the observation that the test might not therefore provide the kind of evidence needed to locate a learner at the C1 level with great confidence. The further the interpretation of the results departs from the aspects of language covered by the test, the more caution must be applied.

Second, there is often some ambiguity in interpreting the CEFR levels and what is meant, for example, by a B1 learner. The definition used by the panellists was of a person with 'the competencies, skills and abilities to be labelled as "B1", but only to such an extent that the slightest decrease in those competencies, skills and abilities would suffice in order not to grant this qualification' (Council of Europe 2009, p.73). This is not necessarily the interpretation used in language schools where a B1 learner is typically a person working towards B1 level objectives. The language school interpretation of ‘B1 learner’ might admit some learners who would be classed as A2 according to the definition provided in the Council of Europe Manual. This suggests that schools may wish to set the entry points for classes working towards CEFR objectives a little lower than in Table 29. Where greater confidence is needed – where it is important that a learner is more than minimally at a certain level – higher cut points may be appropriate.

Ultimately, placement tests must be judged according to their effectiveness at placing students into the most suitable classes. The validity of these recommendations as a basis for local student placement decisions will need to be established through experience. Further studies will be needed to establish that these cut scores meet local needs and that they continue to be appropriate over time. As the body of evidence about the VEPT increases, it will be possible to carry out further studies on the VEPT-CEFR relationship employing person-centred methods. This should also help to establish the extent of agreement concerning the interpretation of the CEFR levels across educational contexts and across continents.

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg: Language Policy Division.

Council of Europe (2009). *Relating examinations to the Common European Framework of Reference: A Manual.* Strasbourg: Language Policy Division.

Godfrey, J.J. & Holliman, E. (1997). *Switchboard-1 Release 2*. LDC Catalog No.: LCD97S62. http://www.ldc.upenn.edu.

Myford, C.M. and Wolfe, E.W. (2003) Detecting and Measuring Rater Effects using Many-Facet Rasch Measurement: Part I*. Journal of Applied Measurement, 4*(4), 386-421

Oakeshott-Taylor, J. (1977). Information redundancy, and listening comprehension. In R. Dirven (ed.), *Hörverständnis im Fremdsprachenunterrict. Listening comprehension in foreign language teaching*. Kronberg/Ts.: Scriptor.

Oller, J. W. (1971). Dictation as a device for testing foreign language proficiency. *English Language Teaching*, 25(3),254-259.

Pearson Education (2011). Versant English Placement Test: Technical Paper. Harlow: Pearson.

Pearson Education (2013). Versant Writing Test: Description and Validation Summary. Harlow: Pearson.

Sigott, G. (2004). *Towards identifying the C-test construct.* New York: Peter Lang.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing, 14*(2), 214-231.

Appendix 1

Table 30. Results of the online familiarisation exercise

CEFR level	***Descriptor***	***A1***	***A2***	***B1***	***B2***	***C1***	***C2***	***Level (assigned by panel)***
A1	Learners can use some words and phrases meaningfully in everyday contexts.	**7**	2	3				A1
A1	Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.	**10**	2					A1
A1	Learners can form simple main clauses but not necessarily accurately.	5	**6**	1				*A2*
A1	Learners know the approximate meaning of around 500 words used in the most common contexts.	3	**6**	3				*A2*
A2	Learners can formulate simple sentences and combine them to some degree by using the most frequent conjunctions, e.g. and, or, because.	1	**9**	2				A2
A2	Learners can use the most frequent verbs in the basic tenses (present and simple past).	2	**8**	2				A2
A2	Learners know the opposites of very frequent words, and can recognize synonyms for some words.	1	**6**	2	3			A2
A2	Learners use some simple structures correctly.	3	**9**					A2
A2	Can link groups of words with simple connectors like 'and', 'but' and 'because'.		**11**		1			A2
A2	Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people.	4	**7**	1				A2
A2	Uses some simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forgets to mark agreement	1	**7**	4				A2
A2	Learners recognize some basic principles of word formation, and can also apply some of them, e.g. to drive — a driver.	2	4	**6**				*B1*
B1	Learners can use some frequent collocations, e.g. a tall girl — a high mountain.	1	**7**	3	1			*A2*
B1	Learners know and can use everyday vocabulary related to a range of basic personal and familiar situations.	2	**6**	4				*A2*
B1	Learners know the basic word order well	2	**6**	3	1			*A2*
B1	Learners know the comparison of adjectives		**8**	4				*A2*
B1	Learners recognize and know how to use the basic word formation principles, e.g. note — notify; fur —furry; accident — accidental; paint — painting.		2	**8**	2			B1
B1	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.		4	**8**				B1
B1	Has sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	1	4	**6**	1			B1
B1	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.		**6**	**6**				B1
B2	Learners can use articles for the most part appropriately.		2	**5**	4	1		*B1*
B2	Learners have a good command of vocabulary related to everyday situations.		1	**9**	2			*B1*
B2	Learners know a number of frequently used idioms.		1	5	3	2	1	*B1*
B2	Learners know the basic principles of constructing passives.		2	9	1			*B1*
B2	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse.		1	8	3			*B1*
B2	Learners can recognize words with more than one meaning e.g. back a car/ back a proposal.		2	1	**8**	1		B2

CEFR level	***Descriptor***	***A1***	***A2***	***B1***	***B2***	***C1***	***C2***	***Level (assigned by panel)***
B2	Learners can use conjunctions in order to express causal and conditional relationships.		1	3	**8**			B2
B2	Learners have a repertoire of frequently used grammatical patterns, which they can use reasonably accurately.		1	4	**6**	1		B2
B2	Can use some complex sentence forms to express viewpoints and develop arguments.			2	**7**	3		B2
B2	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.				**10**	2		B2
C1	Learners can produce complex expressions, such as relative clauses naturally and accurately.			1	**6**	4	1	*B2*
C1	Learners can select an appropriate formulation from a broad range of language to express themselves clearly, without having to restrict what they want to say.				2	**8**	2	C1
C1	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.					**8**	4	C1
C1	Occasional minor slips, but no significant vocabulary errors.				2	**9**	1	C1
C2	Learners can choose structures that fit the topic and style/genre of the text very well.			1	1	**5**	**5**	C2
C2	Learners maintain consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).					4	**8**	C2

Appendix 2

Forms A9 to A22 and A24 of the Manual applied to the VEPT

A3.3 Reception

Listening Comprehension

	Short description and/or reference
1 In what contexts (**domains**, **situations**, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication **themes** are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which **communicative tasks**, **activities** and **strategies** are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Understand and ask questions concerning the tasks, understand instructions, communicate appropriately with others, participate in social life, listening to recordings, listening to overheard conversations, listening for gist, listening for specific information, listening for detailed understanding, listening for implications.
4 What **text-types** and what **length of text** are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: instructions, announcements, interpersonal dialogues and conversations. Minimum text length: 3 words (Repeat). Maximum text length: 48 words (Conversations).
5 After reading the scale for Overall Listening Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for listening comprehension in CEFR 4.4.2.1 listed after the scale might be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no listening comprehension subtest. All tasks are integrative. The tasks that require listening are Repeat, Sentence build, Conversation, and Dictation. The tasks that contribute to the listening subscore Conversation and Dictation. Items at the A1 and A2 level were written specifically to match CEFR “Can do” statements at those levels. Vocabulary for these items was limited to top 1200 most frequent words in our lexicon (which comes from a large corpora of spoken English). Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR. For more evidence see: Versant English Test Validation Report

Form A9: Listening Comprehension

Reading Comprehension

	Short description and/or reference
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle?	Understand and ask questions concerning the tasks to be performed, understand instructions, communicate appropriately

The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	with others, reading for general orientation, reading for information, reading and following instructions, reading for gist, reading for specific information, reading for detailed understanding, reading for implications.
4 What text-types and what length of text are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: computer, print Text types: Short narrative passages, short expository passages with SMOG level between grades 9 and 11, single sentences representing conversational speech. Minimum text length: 4 words (Sentence Completion). Maximum text length: 300 words (Summary and Opinion).

5 After reading the scale for Overall Reading Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for reading comprehension in CEFR 4.4.2.2 listed after the scale might be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no reading comprehension subtest. All tasks are integrative. The tasks that require reading are Read Aloud, Sentence Completion, Passage Reconstruction, and Summary and Opinion. The tasks Read Aloud, Sentence Completion, and Summary and Opinion contribute to the Reading subscore. For all task types except Summary and Opinion, items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR. Summary and Opinion items were written to be at the A2 level. All have a SMOG level between grade 9 and 11. For more evidence see: VEPT Item Specification Document

Form A10: Reading Comprehension

A3.2 Interaction

Spoken Interaction	**Short description and/or reference**
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Understand and ask questions concerning the tasks, understand instructions, communicate appropriately with others, participate in social life, casual conversation, informal discussion.
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: instructions, announcements, interpersonal dialogues and conversations. Minimum text length: 3 words (Repeat). Maximum text length: 48 words (Conversation).
5 After reading the scale for Overall Spoken Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for spoken interaction in CEFR 4.4.3.1 listed after the scale might be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no spoken interaction subtest. All tasks are integrative. The tasks that require speaking are Read Aloud, Repeat, and Sentence build. Each of these tasks contributes to the Speaking subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Form A11: Spoken Interaction

Written Interaction	**Short description and/or reference**
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Communicate appropriately with others, completing sentences, taking down messages from dictation, writing summaries, writing opinions.
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: Single word (sentence completion), single sentence (dictation), summaries of expository and narrative texts, opinions of narrative text. Minimum text length: 1 word (on sentence completion task). Expected minimum 25 words (for Summary). Expected minimum 50 words (for Opinion). Expected minimum 50 words (for Passage Reconstruction). No maximum length, but 18 minute maximum time for longest task (Summary and Opinion).
5 After reading the scale for Overall Written Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for written interaction in CEFR 4.4.3.4 listed after the scale might be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no written interaction subtest. All tasks are integrative. The tasks that require writing are Sentence Completion, Dictation, Passage Reconstruction, and Summary and Opinion. The task types of Passage Reconstruction and Summary and Opinion contribute to the Writing subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Form A12: Written Interaction

A3.3 Production

Spoken Production	**Short description and/or reference**
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Ask questions concerning tasks, communicate appropriately with others, participate in social life, casual conversation, informal discussion.
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: interpersonal dialogues and conversations. Minimum text length: 3 words (Repeat). Maximum text length: 55 words (Read Aloud).
5 After reading the scale for Overall Spoken Production, given	**Level:** A1 to C1

below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for spoken production in CEFR 4.4.1.1 listed after the scale might be of help as a reference.	**Justification (incl. reference to documentation)** There is no spoken production subtest. All tasks are integrative. The tasks that require speaking are Read Aloud, Repeat, and Sentence build. Each of these tasks contributes to the Speaking subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Form A13: Spoken Production

Written Production	**Short description and/or reference**
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Communicate appropriately with others, completing sentences, taking down messages from dictation, writing summaries, writing opinions.
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: Single word (sentence completion), single sentence (dictation), summaries of expository and narrative texts, opinions of narrative text. Minimum text length: 1 word (on sentence completion task). Expected minimum 25 words (for Summary). Expected minimum 50 words (for Opinion). Expected minimum 50 words (for Passage Reconstruction). No maximum length, but 18 minute maximum time for longest task (Summary and Opinion).
5 After reading the scale for Overall Written Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for written production in CEFR 4.4.1.2 listed after the scale might be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no written interaction subtest. All tasks are integrative. The tasks that require writing are Sentence Completion, Dictation, Passage Reconstruction, and Summary and Opinion. The tasks of Passage Reconstruction and Summary and Opinion contribute to the Writing subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Form A14: Written Production

Integrated Skills Combinations		**Task type it occurs in**
1 Listening and Note-taking	□	
2 Listening and Spoken Production	**X**	Repeat, Sentence Completion, Conversation
3 Listening and Written Production	**X**	Dictation
4 Reading and Note-taking	□	
5 Reading and Spoken Production	**X**	Read Aloud
6 Reading and Written Production	**X**	Sentence Completion, Passage Reconstruction, Summary & Opinion
7 Listening and Reading, plus Note-taking	□	
8 Listening and Reading, plus Spoken Production	□	
9 Listening and Reading, plus Written Production	□	

	Complete for each combination
Integrated Skills	**Short description and/or reference**
1 Which skills combinations occur? Refer to your entry in Form A15.	2. Listening and Spoken Production
2 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	Task Type: Repeat Input Text Medium: Spoken Output Text Medium: Spoken Meaning Preserving: Yes Task Types: Sentence build and Conversation Input Text Medium: Spoken Output Text Medium: Spoken Meaning Preserving: No
3 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Understand and ask questions concerning tasks, understand instructions, communicate appropriately with others, participate in social life, casual conversation, informal discussion.
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: instructions, announcements, interpersonal dialogues and conversations. Minimum text length: 3 words (Repeat). Maximum text length: 48 words (Conversation).
7 After reading the scales for Processing Text, given below, plus Comprehension and Written Production given earlier, indicate and justify at which level(s) of the scale the subtest should be situated. The subscale for Note-taking in CEFR 4.6.3 might also be of help as a reference.	**Level:** A1 to C1
	Justification (incl. reference to documentation) There is no "Listening and Spoken Production" subtest. All tasks are integrative. The tasks that require listening are Repeat, Sentence build, Conversation, and Dictation. The tasks that require speaking are Read Aloud, Repeat, and Sentence build. The tasks Conversation and Dictation contribute to the Listening subscores. The task Read Alouds, Repeat, and Sentence build contribute to the Speaking subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Integrated Skills	**Short description and/or reference**
1 Which skills combinations occur? Refer to your entry in Form A15.	3. Listening and Written Production
2 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	Task Types: Dictation Input Text Medium: Spoken Output Text Medium: Written Meaning Preserving: Yes
3 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
4 Which communication themes are the test takers expected to be	Personal identification, house and home, environment, daily life,

able to handle? The lists in CEFR 4.2 might be of help as a reference.	free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Understand instructions, listening to recordings, listening to overheard conversations, listening for gist, listening for specific information, taking down messages from dictation.
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: Voice, computer, print Text type: instructions, announcements, interpersonal dialogues and conversations. Items are single sentence. Minimum text length: 3 words. Maximum text length: 15 words.
7 After reading the scales for Processing Text, given below, plus Comprehension and Written Production given earlier, indicate and justify at which level(s) of the scale the subtest should be situated. The subscale for Note-taking in CEFR 4.6.3 might also be of help as a reference.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** There is no "Listening and Written Production" subtest. All tasks are integrative. The tasks that require listening are Repeat, Sentence build, Conversation, and Dictation. The tasks that require writing are Sentence Completion, Dictation, Passage Reconstruction, and Summary and Opinion. The tasks Conversation and Dictation contribute to the Listening subscore. The tasks Passage Reconstruction and Summary and Opinion, contribute to the Writing subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Integrated Skills	**Short description and/or reference**
1 Which skills combinations occur? Refer to your entry in Form A15.	5. Reading and Spoken Production
2 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	Task Types: Read Alouds Input Text Medium: Written Output Text Medium: Spoken Meaning Preserving: Yes
3 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Reading for general orientation, reading and following instructions.
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: computer, print Text types: Short expository passages with SMOG level between 9 and 11, Minimum text length- to read: 3 words (Sentence Completion). Maximum text length- to read: 300 words. Minimum text length- to speak: 3 words (Repeat). Maximum text length to speak: 55 words (Read Aloud).

7 After reading the scales for Processing Text, given below, plus Comprehension and Written Production given earlier, indicate and justify at which level(s) of the scale the subtest should be situated. The subscale for Note-taking in CEFR 4.6.3 might also be of help as a reference.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** There is no "Reading and Spoken Production" subtest. The tasks that require reading are Read Aloud, Sentence Completion, Passage Reconstruction, and Summary and Opinion. The tasks that require speaking are Read Aloud, Repeat, and Sentence build. The tasks Read Aloud, Sentence Completion, and Summary and Opinion contribute to the Reading subscore. The tasks Read Aloud, Repeat, and Sentence build contribute to the Speaking subscore. Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.

Integrated Skills	**Short description and/or reference**
1 Which skills combinations occur? Refer to your entry in Form A15.	6. Reading and Written Production
2 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	Task Type: Sentence Completion Input Text Medium: Written Output Text Medium: Spoken Meaning Preserving: Yes Task Types: Passage Reconstruction, Summary and Opinion Input Text Medium: Written Output Text Medium: Spoken Meaning Preserving: Yes and No (Passage Reconstruction and Summary involve preserving meaning, but allow for retelling with different vocabulary and grammatical structure. Opinion does not preserve meaning.)
3 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	Domains: Personal, Public, Occupational, Educational
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	Personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather.
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	Communicate appropriately with others, reading for general orientation, reading for information, reading and following instructions, reading for gist, reading for specific information, reading for detailed understanding, reading for implications, writing summaries, writing opinions.
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	Media: computer, print Text types- Input: Short narrative passages, short expository passages with SMOG level between 9 and 11, single sentences representing conversational speech. Minimum text length: 4 words. Maximum text length: 300 words. Text types- writing: Single word (Sentence Completion), summaries of expository and narrative texts, opinions of narrative text. Minimum text length- writing: 1 word (Sentence completion task). (Expected minimum 25 words for Summary and 50 words both for Opinion and for Passage Reconstruction). No maximum length for writing, but 18 minute maximum time for longest task (Summary and Opinion). Text types- reading: Single stand-alone sentences (Sentence completion), Expository and narrative texts (Read Aloud, Passage Reconstruction, Summary and Opinion). Minimum text length- reading: 4 words (Sentence Completion). Maximum text length-

	reading: 300 words (Summary & Opinion).
7 After reading the scales for Processing Text, given below, plus Comprehension and Written Production given earlier, indicate and justify at which level(s) of the scale the subtest should be situated. The subscale for Note-taking in CEFR 4.6.3 might also be of help as a reference.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** There is no "Reading and Written Production" subtest. The tasks that require reading are Read Aloud, Sentence Completion, Passage Reconstruction, and Summary and Opinion. The tasks that require writing are Sentence Completion, Dictation, Passage Reconstruction, and Summary and Opinion. The tasks Read Aloud, Sentence Completion, and Summary and Opinion contribute to the Reading subscore. The task types of Passage Reconstruction and Summary and Opinion contribute to the Writing subscore.

Form A16: Integrated Skills

Linguistic Competence	**Short description and/or reference**
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	Lexical: Fixed expressions, fixed frames, phrasal verbs, compound prepositions, fixed collocations, single word forms. (No proverbs, idioms, archaisms, frozen metaphors.) For all item types except Summary and Opinion, the vocabulary used in test items was restricted to forms of the 8,000 most frequent words found in the Switchboard Corpus (Godfrey and Holliman, 1997), a corpus of three million words taken from spontaneous telephone conversations. In general, the language structures used in the test reflect those that are common in everyday English. This includes extensive use of pronominal expressions such as "she" or "their friend" and contracted forms such as "won't" and "I'm." For Summary and Opinion, the vocabulary used in test items was restricted to the 1600 most frequent words found in the Longman Corpus. Grammatical: all tenses, articles, quantifiers, demonstratives, personal pronouns, question words, relatives, possessives, preposition, auxiliary verbs, conjunctions, particles.
2 After reading the scale for Linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR.
Socio-linguistic Competence	**Short description and/or reference**
3 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	By design, we do not measure socio-linguistic competence. However, results of the VEPT correlate with well with other tests that do measure socio-linguisitc competence.
4 After reading the scale for Socio-linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** n/a **Justification (incl. reference to documentation)** n/a

Form A19: Aspects of Language Competence in Reception

Pragmatic Competence	**Short description and/or reference**
5 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	Knowledge of text design

The lists in CEFR 5.2.3 might be of help as a reference.	
6 After reading the scale for Pragmatic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to B2 **Justification (incl. reference to documentation)** Test items are constructed to measure test taker ability to understand description or narrative, identify main points from relevant supporting detail and examples, and understand detailed information reliably.
Strategic Competence	**Short description and/or reference**
7 What are the strategic competences that the test takers are expected to be able to handle? The discussion in CEFR 4.4.2.4. might be of help as a reference.	Planning, execution, evaluation, and repair (self-correcting and self-editing).
8 After reading the scale for Strategic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** Test items designed to measure skill at using contextual, grammatical and lexical cues to infer intentions and anticipate what will come next.

Form A19: Aspects of Language Competence in Reception

Linguistic Competence	**Short description and/or reference**
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	Lexical: Fixed expressions, fixed frames, phrasal verbs, compound prepositions, fixed collocations, single word forms. All grammatical expected.
2 What is the range of phonological and orthographic competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.4 and 5.2.1.5 might be of help as a reference.	Phonological competence: perception and production of phonemes, distinctive features, syllable structure, prosody, lexical stress patterns, coarticulation. Orthographic competence: knowledge of spelling conventions including contracted forms, knowledge of punctuation, ability to resolve ambiguity, knowledge of the form of letters in printed form in upper case and lower case.
3 After reading the scales for Range and Accuracy in Table A4, indicate and justify at which level(s) of the scale the examination should be situated. The scales for Phonological Control in CEFR 5.2.1.4 and for Orthographic Control in 5.2.1.5 might also be of help as a reference.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR. Scoring reflects test taker's command of a broad lexical repertoire, and command of idiomatic expressions and colloquialisms.
Socio-linguistic Competence	**Short description and/or reference**
4 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	By design, we do not measure socio-linguistic competence. However, results of the VEPT correlate with well with other tests that do measure socio-linguistic competence**.**
5 After reading the scale for Socio-linguistic Competence in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** n/a **Justification (incl. reference to documentation)** n/a
Pragmatic Competence	**Short description and/or reference**
6 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences?	Knowledge of text design, discourse competence

The lists in CEFR 5.2.3 might be of help as a reference.	
7 After reading the scale for Fluency in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1
	Justification (incl. reference to documentation) Scoring measures test taker ability to produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices. Also measures test taker ability to give elaborate descriptions and narratives, develop particular points and round off with an appropriate conclusion.

Form A20: Aspects of Language Competence in Interaction

Strategic Competence	**Short description and/or reference**
8 What are the interaction strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.3.5 might be of help as a reference.	Planning, execution, evaluation, and repair (self-correcting and self-editing).
9 After reading the scale for Interaction in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1
	Justification (incl. reference to documentation) Scoring reflects test taker ability to select a suitable phrase from a readily available range of discourse functions to preface his or her written remarks in order to relate his/her own contributions skilfully to those of a provided text.

Form A20: Aspects of Language Competence in Interaction

Linguistic Competence	**Short description and/or reference**
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	Lexical: Fixed expressions, fixed frames, phrasal verbs, compound prepositions, fixed collocations, single word forms. All grammatical expected.
2 What is the range of phonological and orthographic competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.4 and 5.2.1.5 might be of help as a reference.	Phonological competence: perception and production of phonemes, distinctive features, syllable structure, prosody, lexical stress patterns, coarticulation. Orthographic competence: knowledge of spelling conventions including contracted forms, knowledge of punctuation, ability to resolve ambiguity, knowledge of the form of letters in printed form in upper case and lower case.
3 After reading the scales for Range and Accuracy in Table A5 indicate and justify at which level(s) of the scale the examination should be situated. The scales for Phonological Control in CEFR 5.2.1.4 and for Orthographic Control in 5.2.1.5 might also be of help as a reference.	**Level:** A1 to C1 **Justification (incl. reference to documentation)** Items at the A1 and A2 level were written specifically to match CEFR "Can do" statements at those levels. Items aimed at higher levels (B1 to C1) were drawn primarily from the Versant English Test, whose data driven models demonstrate concordance with the CEFR. Scoring reflects test taker's command of a broad lexical repertoire, and command of idiomatic expressions and colloquialisms.

Form A21: Aspects of Language Competence in Production

Socio-linguistic Competence	**Short description and/or reference**
4 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	By design, we do not measure socio-linguistic competence. However, results of the VEPT correlate with well with other tests that do measure socio-linguistic competence.
5 After reading the scale for Socio-linguistic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** n/a
	Justification (incl. reference to documentation) n/a
Pragmatic Competence	**Short description and/or reference**
6 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	Knowledge of text design, discourse competence
7 After reading the scale for Pragmatic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1
	Justification (incl. reference to documentation) Scoring measures test taker ability to produce clear, smoothly flowing, well-structured text, showing controlled use of organizational patterns, connectors and cohesive devices. Also measures test taker ability to give elaborate descriptions and narratives, develop particular points and round off with an appropriate conclusion.
Strategic Competence	**Short description and/or reference**
8 What are the production strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.1.3 might be of help as a reference.	Planning, execution, evaluation, and repair (self-correcting and self-editing).
9 After reading the scale for Strategic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	**Level:** A1 to C1
	Justification (incl. reference to documentation) Scoring reflects test taker ability to backtrack or revise when he/she encounters a difficulty and reformulate what he/she wants to say without fully altering the flow of speech or written production.

Form A21: Aspects of Language Competence in ProductionSection A5: Specification: Outcome of the Analysis (Chapter 4)

Confirmed Estimation of Overall CEFR Level		
☒ **A1**	☒ **B1**	☒ **C1**
☒	☒	☒
☒ **A2**	☒ **B2**	□ **C2**
☒	☒	□
Short rationale, reference to documentation. If this form presents a different conclusion to the initial estimation in Form A8, please comment on the principal reasons for the revised view.		

Form A24: Confirmed Estimation of Overall Examination Level