Versant™ Professional English Test

Test Description and Validation Summary

Table of Contents

1. Introduction

The Versant™ Professional English Test (VPET), powered by Versant technology, is an assessment instrument designed to measure the ability to understand spoken and written English used in everyday workplace situations, as well as to respond appropriately in diverse spoken and written tasks in an international workplace at an appropriate pace and in intelligible English. The VPET is intended for test-takers over the age of 16 and takes approximately 60 minutes to complete. Because the VPET is delivered automatically by the Versant testing system, the test can be taken at any time, from any location via computer. A human examiner is not required. The computerized scoring provides immediate, objective, and reliable results that correlate well with traditional measures of spoken and written English performance.

VPET scores provide reliable information that can be used for making decisions related to learning and development needs or career advancement by university programs or business and government organizations. The results provide scores, detailed performance summaries, and tips to improve which can be used by academic institutions as well as those who run employability programs and vocational training to monitor progress and measure instructional outcomes. The test results will also provide useful information to test-takers who wish to further improve their English.

2. Test Description

2.1 Test Design

The VPET has two levels (Level 1 and Level 2) corresponding to lower and higher levels of difficulty. Both tests have the same task types and approximately the same number of items, but they differ in their general difficulty. By having two levels, the test becomes more accessible to the test-takers of a wide range of proficiency levels. Also, when a candidate takes a test that is appropriate for their own level, the test results can provide more detailed insights into learners' abilities and gaps. Level 1 is intended for candidates whose English language proficiency is between 10 and 58 on the GSE, which is equivalent to between below A1 and B1+ on the CEFR (Common European Framework of Reference for Languages), an international standard for describing language ability. Level 2 is designed for English learners whose proficiency is between 51 and 90 on GSE, or B1+ to C2 on CEFR.

The VPET has ten tasks: Sentence Completion, Passage Reconstruction, Reading Comprehension, E-mail Writing, Dictation, Response Selection, Passage Comprehension, Repeat, Speaking Situations, and Story Retellings. These tasks measure skills and knowledge that underlie facility in spoken and written English, including pronunciation, fluency, sentence construction, vocabulary, organization, voice and tone, listening, and reading comprehension. Because more than one task contributes to each skill score, the use of multiple tasks strengthens score reliability.

The VPET score report is composed of an Overall score and four skill scores: Speaking, Listening, Reading, and Writing. The Overall score is an average of the four skill scores. These scores indicate the candidate's oral and written proficiency in general and workplace English.

2.2 Test Administration

Tests and test-takers are managed through the web-based administration platform called ScoreKeeper. Because it is taken on a computer, the VPET can be taken at any time, from any location. Automated administration eliminates the need for a human examiner. However, depending on the test score use, a proctor may be needed to verify the candidate's identity and/or to ensure that the test is taken under controlled conditions. To ensure that candidates get the best results, it is recommended that they familiarize themselves with the item types and test format beforehand. An unscored online demo test is available on the Pearson website along with the tips on how to answer spoken questions. If needed, test-takers can also take a full length scored practice test.

Taking the VPET generally takes about 60 minutes. To launch the test, the candidate is required to enter a unique Test Identification Number (TIN). The candidate is then prompted to adjust the volume to an appropriate level and to test the microphone before beginning the test. The test must be taken in a quiet, distraction-free environment, and the candidate must use a microphone headset and speak clearly and naturally. As spoken responses are automatically scored by the machine, having background noise or using an extremely quiet voice may adversely affect test results.

During test administration, an examiner's voice guides the candidate through the test, explains the tasks, and gives examples. The candidate listens through the headset and sees instructions and examples on the computer screen. Candidates respond to test questions by speaking into the microphone, typing on the computer keyboard, or clicking the mouse.

The delivery of some of the item prompts is interactive—the system detects when the candidate has finished responding to an item, and then presents the next item. For other item prompts, the candidate has a set amount of time to respond to each item. If the candidate does not finish a response in the allotted time, whatever response was made is saved automatically and the candidate proceeds to the next item. If candidates finish before the allotted time has run out, they can click a button labeled "Next" to move on to the next item – however, this option is not available for every item type.

When the test is finished, the candidate clicks a button labeled "Finish." The candidate's responses are sent to a remote server where the Versant testing system automatically analyzes them, calculates scores, and posts them to ScoreKeeper, the password-protected test administration platform, usually within minutes of completing the test. Test administrators may choose to share results with test-takers (either on- or offline).

2.3 Number of Items

The VPET is a four-skill English test and the maximum test time is 60 minutes. During a test administration, approximately 58 items are presented to each candidate in the ten separate sections - Parts A through J. The items are drawn at random from a large item pool. This means that most or all items are different from one test administration to the next. Proprietary algorithms are used by the testing system to select from the item pool – the algorithms take into consideration, among other things, an item's difficulty level and similarity to other presented items. Table 1 shows the approximate number of items presented in each section. The exact number of items in each test may change from time to

time as new, unscored items are added to and removed from the test. The responses to the unscored items do not impact the candidates' scores nor do they impact the test experience. The responses are used to build scoring models for new items, which allows Pearson to add new content to the test in order to keep the item bank secure and up-to-date.

Table 1. Approximate number of items presented per section

Part	Task	Approximate Number of Items Presented
A	Sentence Completion	10
B	Passage Reconstruction	3
C	Reading Comprehension	6 (3 passages, 2 questions each)
D	E-mail writing	2
E	Dictation	8
F	Response Selection	8
G	Passage Comprehension	6 (2 passages, 3 questions each)
H	Repeat	10
I	Speaking Situations	2
J	Story Retellings	3
	Total	58

2.4 Test Format

The following subsections provide brief descriptions of the tasks and the abilities required to respond to the items in each of the ten parts of the VPET.

Speech Sample

In this task, candidates listen to a spoken question that asks them to describe something or give their opinion on a topic. Candidates have up to 30 seconds to respond to the question.

Examples:

> Do you prefer speaking with someone by a voice call or a video call? Explain why.
>
> Do you think it's important to learn English? Why or why not?

This task is used to collect a longer spontaneous speech sample. Candidates' responses to items in this section are not scored but are available for review by authorized listeners. These questions are not considered test items.

Part A: Sentence Completion

In this task, candidates read a sentence that has a word missing, and they supply an appropriate word to complete the sentence. Occasionally, two adjacent sentences are presented but still only one word is missing. Candidates are given 25 seconds for each item. During this time, candidates must read and understand the sentence, identify a lexical item to complete the sentence, and type the word above the line provided. Sentences range in length from 7 to 28 words. Across all items in this task, candidates are exposed to sentences with words missing of various parts of speech (e.g., noun, verb, adjective, adverb) and with different positions in sentences: sentence-initial, sentence-medial, sentence-final.

Examples:

1. He always kept a flashlight in his car in case of an _____.
2. She was the most creative problem-solver I ever met; she was never at a _____ for a good idea.

It is sometimes considered that fill-in-the-gap tasks are more authentic when longer passages or paragraphs are presented to the candidate (as in the case of cloze tasks), because this enables context inference strategies. However, research has shown that candidates 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 candidates 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. Therefore sentence-level fill-in-the-gap tasks are appropriate and provide more information than those of longer cloze-type passages.

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 (which can include more than one answer choice when appropriate).

The sentence completion task draws on interpretation, inference, lexical selection and morphological encoding, and as such measures both the reading and writing ability of a candidate.

Part B: Passage Reconstruction

Passage Reconstruction is similar to a task known as free-recall, or immediate-recall, in which candidates 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 candidate has 90 seconds to reconstruct the content of the passage in writing. Passages range in length from 30 to 87 words. The items sample a range of sentence lengths, syntactic variation, and complexity. Two discourse genres are presented in this task: narrative and e-mail. Narrative texts are short stories about common situations involving characters, actions, events, reasons, consequences, or results. E-mail texts are adapted from authentic electronic communication and may be conversational messages to colleagues or more formal messages to customers.

Examples:

> *(Narrative)* Corey is a taxi driver. It is his dream job because he loves driving cars. He started the job ten years ago and has been saving up money since then. Soon, he will use this money to start his own taxi company.
>
> *(E-Mail)* Thank you so much for being so understanding about our delay of shipment. It has been quite difficult to get materials from our suppliers due to the recent weather conditions. It is an unusual circumstance. In any case, we should be able to ship the products to you tomorrow. In the meantime, if you have any questions, please feel free to contact me.

In order to perform this task, the candidate must read the passage presented, understand the concepts and details, and hold them in memory in order to reconstruct the passage. Individual candidates may naturally employ different strategies when performing the task. Reconstruction may be somewhat verbatim in some cases, especially for shorter passages answered by advanced candidates. For longer texts, reconstruction may be accomplished by paraphrasing and drawing on the candidate's own choice of words. Regardless of strategy, the end result is evaluated based on the candidate'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 documenting events or decisions, summarizing documents, or writing the minutes of meetings.

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 candidate'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), it is more an integrated test of both reading and writing (Alderson, 2000).

Part C: Reading Comprehension

In this task, candidates are presented with a passage and two comprehension questions with multiple choice options. The candidate has three minutes to read a passage and answer two questions. The passage consists of written material drawn from everyday workplace situations. The passage and

options may include graphs or charts, but such figures are very basic and require only a limited amount of graphic literacy to understand.

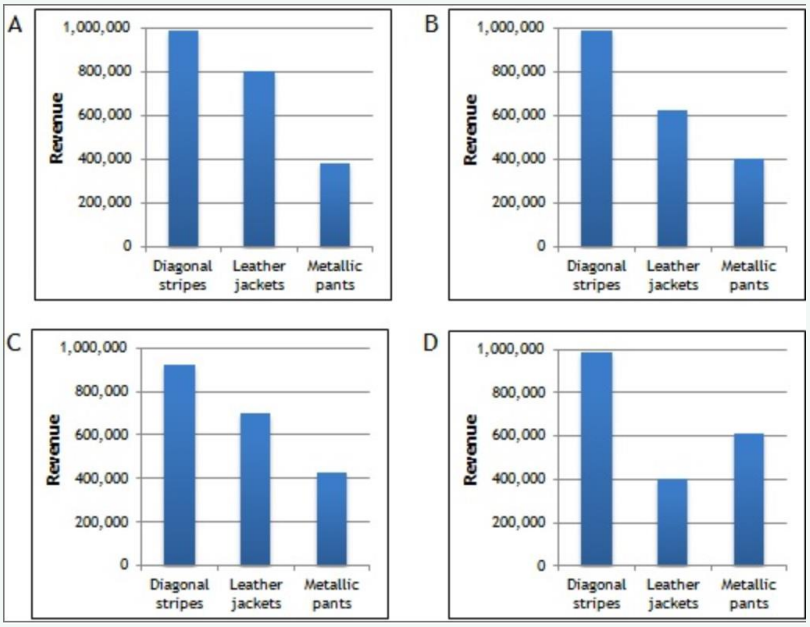
Example:

From an early age, Sally Kwon, “the woman who dresses the nation,” was experimenting with fashion. In high school, she taught herself how to sew. "I always made my own clothes," she says. "I knew what I liked.” She regularly made her own dresses for parties. Now, as a famous designer for Hendrick’s, a global retail store, Kwon has been influencing the fashion industry for years. Kwon took over as the head designer for the chain in 2007. Previously serving as a low level buyer at the company, she quickly demonstrated that she had an eye and passion for fashion. Within two years of becoming a full-time designer, she had made her mark – with the entire fashion industry taking notice. Her selections helped to shape purchasing decisions across the company. In 2009, her daywear line featuring diagonal stripes generated nearly one million dollars in revenue alone. Bold leather jackets represented about six hundred thousand in revenue, and metallic pants generated four hundred thousand in revenue that same year. Given her fashion sense and enthusiasm for setting trends, if you’re not wearing Sally Kwon’s styles yet, you can bet that you will be soon. Want to read more of this interview? Go to: CareersoftheFuture.com/fash-design2

Question 1. What is the best title for the article?

A. A decade of change at Hendricks
B. Buying trends in fashion
C. A fashion star on the rise
D. Declines in fashion trends

Question 2. Which of the following charts best represents the information in the article?



The "passage" portion of Reading Comprehension items may only include one main portion of text, while some passages may contain two portions, which could come from different sources. For example, one text might be an invoice for services rendered while a second could be a letter from a customer disputing a charge. Such passages require integration and synthesis across the two types of text in order to answer questions correctly.
The comprehension questions conform to the following "types", based on which reading skills are required to answer:

- **Main idea**: identify the central theme of the passage
- **Organization**: identify portions of the passage which do not conform to appropriate or logical organization
- **Fact**: locate or verify a particular detail which is explicitly expressed in the text
- **Inference**: answer a question which does not have an explicit referent in the passage

Part D: E-mail Writing

In this task, candidates are given an opportunity to demonstrate their writing ability using e-mail in relatively formal, work/business-related settings. Candidates are presented with a short description of a situation and must write an email in response to the situation. Possible functions which candidates might encounter include, but are not limited to: giving suggestions, making recommendations, requesting information, negotiating a problem, giving feedback, and reporting an event. Candidates are given 9 minutes to read and respond to the situation. Responses of at least 100 words are expected, and those that are less than 30 words or that are off-topic are assigned the lowest possible score.

Each email situation contains several elements:

- the setting or place of work where the correspondence takes place
- the addressee to whom the email is to be written, and the relationship between the candidate and the addressee
- the goal or functional purpose of the email
- three themes (e.g., suggestions, reasons, or recommendations) which the candidate should address in his/her response

Examples:

> The company you work for recently hired several new employees. Your sales team has been asked for ideas about the training they should receive. Write an e-mail to your manager, Mr. Jenkins, suggesting three areas new-employee training should focus on.
>
> Your suggestions must come from the following three themes:
> - company history
> - product knowledge
> - communicating with customers
>
> You should include all three themes. Provide supporting ideas for each of your suggestions.

Candidates are not expected to generate original content for their responses as the themes to address are provided for them. However, candidates are required to construct elaborations, supporting ideas or reasons for each of the themes. In order to fulfill the task, candidates must understand the situation presented, relate it to their existing knowledge, and synthesize and evaluate the information such that an appropriate response can be composed. Candidates must be conscious of the purpose of the email, address each of the themes, and understand the relationship between themselves as the writer and the intended recipient of the email. Candidates must fully understand the prompt in order to construct an informative, organized, succinct response with appropriate tone, word choice, and grammatical accuracy.

Part E: Dictation

In the Dictation task, candidates listen to a sentence and then type it exactly as they hear it. Candidates have 25 seconds to type each sentence. The sentences are presented in approximate order of increasing difficulty. Sentences range in length from 4 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.

Examples:

1. There's hardly any paper left.
2. Corporations and companies are staying current with the latest technologies.

Dictation requires the candidate 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, candidates 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).

The task provides information on comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for candidates 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 F: Response Selection

In the Response Selection task, candidates listen to a sentence, which is immediately followed by three possible responses. From among the three possible responses, candidates choose the one that is the

most appropriate response to the sentence. Candidates answer each question either by clicking 'A', 'B', or 'C'. They are given 8 seconds to respond.

Example:

Didn't I see you at the fundraiser this weekend?

A. No, I haven't seen it, but I heard it was good.
B. Maybe; I might have time to volunteer this weekend.
C. No, it couldn't have been me. I was out of town.

The sentences and possible responses are spoken at a conversational pace. This task is designed to measure candidates' listening comprehension ability. The task demands immediate word recognition and extraction of meaning in the stream of speech, comprehension of the key proposition in the sentence and identification of which response is the best match given the sentential context.

Part G: Passage Comprehension

In the Passage Comprehension task, candidates listen to a spoken passage (usually a story) and then are presented with three comprehension questions about the passage. The passages range from 30 to 87 words in length. Most passages are simple stories with a situation involving a character (or characters), a setting, and an ending. The body of the story typically describes an action performed by the agent of the story followed by a possible reaction or implicit sequence of events. The ending typically introduces a result, new situation, actor, thought, or emotion.

Example:

Jason woke up feeling sick. He called his boss and explained that he could not come in to work. Immediately after making the phone call, he took some medicine. A few hours later, Jason no longer felt sick. Rather than waste the afternoon at home, he decided to go to work after all.

After listening to a passage, the candidate hears and responds to three comprehension questions.

Question 1: What problem did Jason have when he woke up?
Question 2: What did he do right after calling his boss?
Question 3: What did Jason do that afternoon?

For each passage, candidates are asked to answer three comprehension questions. Correct answers to the questions (or information needed for simple inferences) are all included in the passage. Questions typically ask for the main idea and details of the passage. Unlike the Response Selection section, the Passage Comprehension task allows for the assessment of candidates' listening comprehension ability with longer speech.

Part H: Repeat

In this task, candidates are asked to repeat sentences that they hear verbatim. The sentences are presented to the candidate in approximate order of increasing difficulty. Sentences range in length from 4 to 17 words. The audio item prompts are spoken in a conversational manner.

Example:

1. I need to go back to work.
2. The repair person will be here some time this afternoon.

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) rather than simply parrot the phonemes that appeared. Highly proficient speakers of English can generally repeat sentences that contain many more than seven syllables because they 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 candidates to organize speech into linguistic units or chunks, these items assess the candidate's mastery of phrase and sentence structure. Given that the task requires the candidate to repeat full sentences (as opposed to just words and phrases), it also offers a sample of the candidate's fluency and pronunciation in continuous spoken English.

Part I: Speaking Situations

In this task, candidates listen to and read a brief scenario and are then asked to respond as if they were in the situation. Candidates have 10 seconds to prepare a response and 60 seconds to respond to each situation. Candidates are expected to give pragmatically appropriate responses as well as respond using accurate grammar and appropriate connectors and cohesive devises.

Example:

You borrowed a jacket from your friend, Mark. However, you spilled coffee on it, and it left a large stain. Mark calls and says he needs his jacket. What would you say to him?

The Speaking Situations task elicits aspects of pragmatic ability in a relatively open, long turn response. Candidates must demonstrate awareness and appropriate use of the kind of language required in different social situations eliciting speech acts such as apologizing, requesting, and refusing. Responses are scored based on the appropriateness and clarity of the response for the given situation, the effectiveness and extent to which the social demand is conveyed, and the extent to which the candidate used appropriate politeness conventions and spoken register.

Part J: Story Retellings

In this task, candidates listen to a brief story and are then asked to describe what happened in their own words. Candidates have 30 seconds to respond to each story. Candidates are encouraged to tell as much of the story as they can, including the situation, characters, actions and ending. The stories consist of three to six sentences and contain from 30 to 90 words. The situation involves a character (or characters), setting, and goal. The body of the story describes an action by the agent of the story followed by a possible reaction or implicit sequence of events. The ending typically introduces a new situation, actor, thought, or emotion.

Example:

> Paul planned on taking the late flight out of the city. He wasn't sure whether it would be possible because it was snowing quite hard. In the end, the flight was cancelled because there was ice on the runway.

The Story Retellings items assess a candidate's ability to listen and understand a passage, reformulate the passage using his or her own vocabulary and sentence structure, and then retell it in detail. This section elicits longer, more open-ended speech samples than earlier sections in the test and allows for the assessment of a wide range of spoken abilities.

2.5 Test Construct

2.5.1 Workplace Emphasis

VPET is designed to measure a candidate's ability to understand and use English in workplace contexts. The test does not target language use in any specific industry (e.g., banking, accounting, travel, health care) or job category (e.g., shop clerks, accountant, tour guide, nurse), because assessing the candidate's English ability in such specific domains requires both English ability and content knowledge, such as subject matter knowledge or job-specific terminology. Rather, VPET is intended to assess how well and how efficiently the candidate can process and produce English on general topics such as scheduling, commuting, and communicating amongst co-workers, which are commonly found in the workplace regardless of industry or job category.

2.5.2 Facility in Spoken and Written English

For any language test, it is essential to define the test construct as explicitly as possible (Bachman, 1990; Bachman & Palmer, 1996). VPET is designed to measure a candidate's facility in English in the workplace context, which is how well the candidate can understand spoken and written English as well as respond appropriately in diverse spoken and written tasks in an international working environment at an appropriate pace and in intelligible English.

The first concept embodied in the definition of facility is *how well a candidate understands spoken and written English*. Both receptive modalities (listening and reading) are used in the test. Repeat, Story Retellings, Speaking Situations, Response Selection, Passage Comprehension, and Dictation expose candidates to spoken English, and Sentence Completion, Passage Reconstruction, E-mail Writing,

Reading Comprehension, and Speaking Situations present written English that candidates must read and comprehend within given time limits.

Repeat, Story Retellings, Speaking Situations, Response Selection, Passage Comprehension, and Dictation require segmenting the acoustic stream into discrete lexical items and receptively processing spoken language forms including morphology, phrase structure and syntax in real time.

Sentence Completion, Passage Reconstruction, E-mail Writing, Reading Comprehension, and Speaking Situations require fluent word recognition and problem-solving comprehension abilities (Carver, 1991). Interestingly, the initial and simplest step in the reading process—word recognition— is something that differentiates first language readers from even highly proficient second-language readers (Segalowitz, Poulsen, & Komoda, 1991). First language readers have massively over-learned words by encountering them in thousands of contexts, which means that they can access meanings automatically while also anticipating frequently occurring surrounding words.

Proficient language users consume fewer cognitive resources when processing spoken or written language than users of lower proficiency, and they therefore have capacity available for other higher-level comprehension processes. Comprehension is conceived as parsing sentences, making inferences, resolving ambiguities, and integrating new information with existing knowledge (Gough, Ehri, & Trieman, 1992).

The second concept in the definition of facility in spoken and written English is *how well the candidate can respond appropriately in speaking and writing*. The speaking tasks in the VPET are designed to tap into the many kinds of processing required to participate in a spoken conversation: a person has to track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 1, adapted from Levelt (1989).

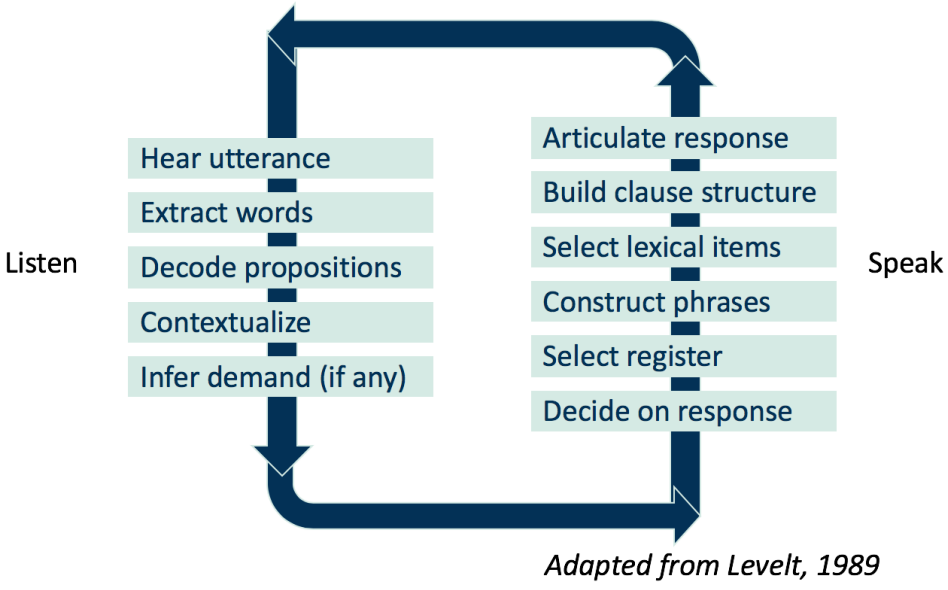


Figure 1. Conversational processing components in listening and speaking

Core language component processes, such as lexical access and syntactic encoding, typically take place at a very rapid pace. Van Turennout, Hagoort, and Brown (1998) found that during spoken conversations, speakers go from building a clause structure to phonetic encoding in about 40 milliseconds. Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a speaker involved in interactive spoken communication. A typical window in turn taking is about 500 to 1000 milliseconds (Bull & Aylett, 1998). If language users cannot perform the internal activities presented in Figure 1 in real time, they will not be able to participate as effective listener/speakers. Thus, spoken language facility is essential in successful oral communication.

The extended writing tasks, Passage Reconstruction and E-mail Writing, are designed to assess not only proficiency in the core linguistic skills of grammatical and lexical range and accuracy, as described above, but also the other essential elements of good writing such as organization and effective expression of ideas. These are not solely language skills but are more associated with effective writing and critical thinking and must be learned. Assuming these skills have been mastered in the writer's first language, they may be transferable and applied in the writer's second language, if their core linguistic skills in second language are sufficiently advanced. Skill in organization may be demonstrated by presenting information in a logical sequence of ideas; highlighting salient points with discourse markers; signposting when introducing new ideas; or giving main ideas before supporting them with details.

The last concept in the definition of facility in spoken and written English is the candidate's ability to perform the requested tasks *at an appropriate pace* in intelligible English. The rate at which a candidate can process spoken language, read fluently, and appropriately respond in speaking and writing plays a critical role in whether or not that individual can successfully communicate in real-world situations. A strict time limit imposed on each item ensures that proficient language users are advantaged and allows the possibility of identifying candidates with different levels of automaticity in this skill.

Automaticity in language processing is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, & Schriefers, 2003; Levelt, 2001). Automaticity is required for the listener/speaker to be able to focus on what needs to be said rather than to how the language code is structured or analyzed. By measuring basic encoding and decoding of oral language as performed in integrated tasks in real time, the VPET probes the degree of automaticity in language performance.

By utilizing integrated tasks, the VPET taps into core linguistic skills and measures the ability to understand and respond to spoken and written English. After initial identification of a word, either as acoustic signal or textual form, candidates who are proficient in the language move on to higher-level prediction and monitoring processes such as *anticipation*. Anticipation enables faster and more accurate decoding of language input, and also underlies a candidate's ability to select appropriate words when producing spoken or written English. The key skill of anticipation is assessed in the Repeat, Dictation, Sentence Completion, and Passage Reconstruction tasks of the VPET as candidates are asked to anticipate missing words and reconstruct texts.

2.5.3 The Role of Memory

Some measures of automaticity can be misconstrued as memory tests. Because some VPET tasks involve repeating long sentences or holding sentences in memory in order to type them, it may seem that these tasks measure memory instead of language ability, or at least that performance on some tasks may be unduly influenced by general memory performance. During the development of the test, every Repeat and Dictation item was presented to a sample of educated native speakers of English, and at least 90% of the speakers in that sample responded correctly. If memory, as such, were an important component of performance on these tasks, then the native English speakers should show greater performance variation on these items according to the presumed range of individuals' memory span.

2.5.4 The Role of Context

The VPET probes the psycholinguistic elements of spoken and written language performance rather than the social, rhetorical, and cognitive elements of communication. In general, VPET items present context-independent material in English. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based (Perry, 2001). Second, when language usage is relatively context-independent, task performance depends less on factors such as world knowledge and cognitive style and more on the candidate's facility with the language itself. Thus, the test performance relates most closely to language abilities and is not confounded with other candidate characteristics. Third, context-independent tasks maximize response density; that is, within the time allotted for the test, the candidate has more time to demonstrate performance in writing the language because less time is spent presenting contexts that situate a language sample or set up a task demand.

The two exceptions to this context-independence are the following tasks: Speaking Situations and E-mail Writing. The E-mail Writing task presents a situation with schema that candidates must attune to, for example, the purpose of the writing and the relationship between themselves and the intended recipient of the e-mail. In this way, E-mail Writing allows for the assessment of the grammar and mechanics of writing, as well as knowledge of the e-mail genre and the rhetorical and cultural norms for organizing information in e-mails. The Speaking Situations task similarly presents a situation requiring candidates to infer the sociolinguistic demand and produce an appropriate response that would successfully satisfy the demand. In both cases, candidates are provided with contextual information to allow them to build schema. Similarly, for both of these tasks, achieving a high score requires the ability to employ sociolinguistic knowledge (i.e., in this situation, what kinds of language, linguistic structures, and tone are appropriate?) and to convey the appropriate sociocultural message using spoken (Speaking Situations) or written (E-mail Writing) English.

3. Content Design and Development

3.1 Vocabulary Selection

The vocabulary used in the test items was taken from a general English corpus and a business English word list. The general English corpus was restricted to forms of the 8,000 most frequent words found in

the Switchboard Corpus (Godfrey & Holliman, 1997), a corpus of three million words taken from spontaneous telephone conversations. The business English word list was restricted to forms of the 3,500 most frequent words found in the University of Cambridge Business English Certificate Preliminary Wordlist, Barron's 600 Essential Words for the TOEIC, and Oxford Business and Finance words.

3.2 Item Development

VPET items were drafted by trained item writers who had advanced degrees or training in applied linguistics, TESOL, or language testing. In general, structures used in the test reflect those that are used in everyday workplace situations. The items employ a wide range of topics from relatively general English domains to common workplace domains. The item writers were provided a list of potential topics/activities/situations with regard to the business domain, such as:

- Announcements
- Business trips
- Complaints
- Customer service
- Phone call, E-Mail
- Inventory
- Scheduling
- Marketing/Sales

Item writers were specifically requested to write items so that items are not specific to a certain business domain and would not favor candidates with work experience or require any work experience to answer correctly. The items are intended to be within the realm of familiarity of both a typical, educated, native English speaker and an educated adult who has never lived in an English-speaking country.

Draft items were then reviewed internally by a team of test developers, all with advanced degrees in language-related fields, to ensure that they conformed to item specifications and English usage in different English-speaking regions and contained appropriate content. Then, draft items were sent to external experts in the U.S., the U.K., and Australia. The pool of expert reviewers included several individuals with PhDs in applied linguists and subject matter experts who worked as training and recruitment managers for large corporations. Expert review was conducted to ensure 1) compliance with the vocabulary specification, and 2) conformity with current colloquial English usage in different countries. Reviewers checked that items would be appropriate for candidates trained to standards other than American English.

All items, including anticipated responses for Sentence Completion, were checked for compliance with the vocabulary specification. Most vocabulary items that were not present in the lexicon were changed to other lexical items that were in the corpus and word list. Some off-list words were kept and added to a supplementary vocabulary list, as deemed necessary and appropriate. The changes proposed by the different reviewers were then reconciled and the original items were edited accordingly.

For an item to be retained in the test, it had to be understood and responded to appropriately by at least 85% of a reference sample of educated native speakers of English.

3.3 Item Prompt Recording

3.3.1 Voice Distribution

19 native speakers (9 women and 10 men) representing various speaking styles and regions, including the U.S. and Australia, were selected for recording the spoken prompt materials. Recordings were made in a professional recording studio in Menlo Park, California. In addition to the item prompt recordings, all the test instructions and listening comprehension questions were also recorded by professional voice talents whose voices were distinct from the item voices.

3.3.2 Recording Review

Multiple independent reviews were performed by test developers on all the recordings for quality, clarity, and conformity to natural conversational styles. Any recording in which reviewers noted some type of error was either re-recorded or excluded from installation in the operational test.

4. Score Reporting

4.1 Scores and Weights

The VPET score report is made up of an Overall score and four skill scores (Speaking, Listening, Reading, and Writing).

Overall: The Overall score on this test reflects a candidate's ability to understand spoken and written English in the international workplace. To get a high score, candidates need to also be able to respond appropriately in various spoken and written tasks. Speaking at a conversational pace and in intelligible English are also important criteria. Overall scores are based on an equally weighted combination of speaking, listening, writing, and reading scores. However, overall scores are withheld if one or more skill scores is BL (Below Level), or if the speaking score is unavailable due to unrecognizable speech.

Speaking: The Speaking score reflects the ability to communicate in a range of workplace situations, including business and social interactions. The score is based on the ability to produce fluent, intelligible speech by using appropriate stress, rhythm, and intonation as well as accurate grammar. In the case of the Level 2 test, a speaking score below the range of a test level (i.e., lower than 51) is reported as BL (Below Level). Also, in cases where there are not sufficient spoken responses or the speech is hard to recognize due to very soft voice, whispering, or background sounds, the speaking score may not be available.

Listening: The Listening score reflects the ability to understand main ideas and specific details from a range of everyday workplace speech. The score is based on the ability to comprehend the meaning of English spoken at a normal conversational speed. In case of Level 2 test, a listening score below the range of a test level (i.e., lower than 51) is reported as BL (Below Level).

Writing: The Writing score reflects the ability to produce a variety of texts on everyday workplace topics. The score is based on the ability of the candidate to express him/herself with clear, effective

structure as well as appropriate tone and style according to the purpose and audience of the text. In case of Level 2 test, a writing score below the range of a test level (i.e., lower than 51) is reported as BL (Below Level).

Reading: The Reading score reflects the ability to understand written English texts on everyday workplace topics. The score is based on the ability to operate at a functional speed to extract details and main ideas, infer the message, and construct meaning. In case of Level 2 test, a reading score below the range of a test level (i.e., lower than 51) is reported as BL (Below Level).

Table 2 shows how the four skill scores are weighted to achieve an Overall score.

Table 2. Skill score weighting in relation to VPET Overall score

Skill Score	Weight
Speaking	25%
Listening	25%
Reading	25%
Writing	25%
Overall	**100%**

In the VPET scoring logic, the four skill scores are weighted equally because successful communication depends on all four skills. Producing accurate spoken and written content is important, but poor listening or reading comprehension skills can lead to inappropriate responses; in the same way, accurate listening and reading comprehension skills without the ability to articulate or write an appropriate response can also hinder communication. This is why an overall score is withheld in case one or more skill scores are BL (Below Level) or the speaking score is unavailable.

Each incoming spoken response from a VPET is recognized automatically by a speech recognizer that has been optimized for non-native speech. The words, pauses, syllables, phones, and even some subphonemic events are located in the recorded signal. The content of the responses to Repeat, Speaking Situations, and Story Retellings is scored according to the presence or absence of expected correct words in correct sequences. The manner of the response (fluency and pronunciation) is calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. These measures are scaled according to the native and non-native distributions and then re-scaled and combined so that they optimally predict human judgments on manner of speaking.

Each incoming written response from a VPET is recognized automatically by the Versant testing system. The content of the responses to Sentence Completion and Dictation are scored according to the presence or absence of expected correct words in correct sequences. The content of responses to Passage Reconstruction and E-mail Writing items are scored for content by scaling the weighted sum of the occurrence of a large set of expected words and word sequences in the written response. Weights are assigned to the expected words and word sequences according to their semantic relation to the prompt using a variation of latent semantic analysis (Landauer, Foltz, & Laham, 1998). These responses are also scored for grammar, spelling, punctuation, capitalization, and syntax.

4.2 Global Scale of English

Test scores on VPET are reported on the Global Scale of English that ranges from 10 to 90. It is designed to report English language proficiency accurately and easily. It has also been empirically linked with and enhances the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) by showing finer gradations of a learner's level and progress within a CEFR band, as shown in Figure 3.

Figure 3. GSE and CEFR Mapping

Table 3 shows the mapping of GSE score ranges with the CEFR levels.

Table 3. Mapping of CEFR Levels with GSE scores

CEFR Level	GSE Score Range
<A1	10-21
A1	22-29
A2	30-42
B1	43-58
B2	59-75
C1	76-84
C2	85-90

GSE is an ecosystem that consists of the scale itself, the GSE Learning Objectives, course materials and assessment. Extensive documentation about the scale and how it can be used is available at https://www.pearson.com/english/about/gse.html.

4.3 Score Use

Once a candidate has completed a test, the candidate's responses are sent to a remote server, from which the Versant testing system analyzes and scores them. Test results are available on ScoreKeeper, the password-protected test administration platform, and the test administrators can choose to share the results with the candidates on or offline. Score users of VPET may be business organizations, educational and government institutions. Pearson endorses the use of test results for making decisions about the English skills of individuals, provided score users have reliable evidence confirming the identity

of the individuals at the time of test administration. Score users may obtain such evidence either by administering the VPET themselves under secure conditions, or by having trusted third parties administer the test.

VPET scores can be used to assess how well and efficiently a candidate can process and produce spoken and written English in the workplace and professional setting. Business organizations may use VPET scores as part of the screening, hiring, language monitoring, or promotion process. Educational institutions and language programs can use VPET scores to evaluate the level of English proficiency of individuals entering into, progressing through, and leaving business English language courses. Individual test-takers can use the information provided in the score report such as descriptors of their current level of English and recommendations on how to further improve their English in order to plan their own learning. The VPET scores cover a wide range of abilities in spoken and written English communication. In most cases, score users must decide what score is considered a minimum requirement in their context (i.e., a cut score). Score users may wish to base their selection of an appropriate cut score on their own localized research. Pearson can provide assistance in helping organizations to arrive at data-based cut scores.

5. Field Testing

As the VPET has been developed using the test items included in PTE Professional, a separate field test was not carried out for the VPET for the following purposes: (1) to check the performance of the test items with both native speakers and learners, (2) to calibrate the difficulty of each item based on a large sample of candidates at various proficiency levels and from various first language backgrounds, and (3) to collect sufficient written and spoken English samples to develop automatic scoring models for the test. Therefore, Sections 5 and 6 are based on what was reported for PTE-Pro (Pearson, 2017).

To field test PTE-Pro items, both native speakers of English and English language learners were recruited as participants to take a data collection version of the items. Candidates recruited for the field testing took one or more of three forms of PTE Professional, which included: (1) a modified form which only had Speaking and Listening items (i.e., PTE Professional Speaking Profile); (2) a modified form which only had Writing and Reading items (i.e., PTE Professional Writing Profile); or (3) a combination of both modified forms.

5.1 Native Speakers

A total of 73 educated adult native English speakers were recruited. Most were from the U.S. with a few from the U.K. and Australia. Most native English speakers took the test multiple times producing a total of 706 completed tests. Each test was comprised of a unique set of items, so items did not overlap between the tests. The mean age of the native speaker sample was 35.6 and the male to female ratio was 31% to 63% with the remaining 6% not reporting their gender.

While PTE Professional was specifically designed for English learners, responses from native speakers were used to validate the appropriateness of the test items and their performance was also used to evaluate the scoring models.

5.2 English Language Learners

For the Writing Profile version of PTE Professional, a total of 1,695 English language learners were recruited from various countries representing both university students and working professionals. A total of 46 countries were represented in the field test, but the majority of the data were collected in Argentina, China, Germany, India, Italy, Japan, Korea, Philippines, Spain, and Taiwan. A total of 55 different languages were reported. The male to female proportion was 44% to 49%, with the remaining 7% of the candidates not reporting their gender. The mean candidate age was 28 years.

For the Speaking Profile version of PTE Professional, a total of 973 English learners were recruited from various countries representing both university students and working professionals. All the tests conducted for the purpose of data collection were taken on the computer. Almost all of the candidates took the test only once. Based on the country of origin, there were a total of 46 countries represented in the field test. Similar to the field study for the PTE Professional Writing Profile, the majority of the data collected for the Speaking Profile were from Argentina, China, Germany, India, Italy, Japan, Korea, the Philippines, Spain, and Taiwan. A total of 46 different languages were reported. The male to female proportion was 50% to 47%, with the remaining 3% of the candidates not reporting their gender. The mean age was 28.9 years.

6. Data Resources for Scoring Development

As a result of the field test, more than 200,000 spoken responses and 50,000 written responses were collected from native speakers of English and English learners. The response data were stored in a database to which the test development experts had access for various purposes such as transcribing, rating, and development of scoring models. A particularly resource-intensive undertaking involved transcribing spoken responses: the vast majority of native speaker responses were transcribed at least once, and the majority of the English learner responses were transcribed two or more times to ensure that the most accurate transcriptions were used to build the scoring models.

6.1 Transcription

A subset of the spoken responses was transcribed by a team of trained transcribers. The purpose of transcribing spoken responses was to transform the audio recorded responses into annotated orthographic text, and to then use the transcriptions to develop and validate automated scoring systems based on a large sample of candidates at various levels and from various first language backgrounds.

Responses were transcribed by a group of educated native speakers of English located in the United States. They all underwent rigorous training, which included understanding the purpose of transcription

and learning a specific set of rules and annotation symbols. Subsequently, they completed a series of training sets. Only the transcribers meeting the standards on the training sets were selected. During the actual transcription process, the quality of the transcriptions was closely monitored by the test development team, and the transcribers were given feedback throughout the process. As an additional quality check, when two transcriptions for the same response did not match, the response was automatically sent to a third transcriber for adjudication.

The actual transcription process was carried out using an online interface. Transcribers could listen to each response as many times as they wished in order to understand the response. Audio was presented from a stack semi-randomly, so that a candidate's set of responses would be spread among many different transcribers.

A total of 30,718 transcriptions were produced for native responses, and 87,746 transcriptions were produced for learner responses.

6.2 Human Rating

Field test responses were rated by expert raters according to a set of rubrics in order to provide criteria for the machine scoring. That is, machine scores were developed to predict the expert human scores. Selected item responses to Passage Reconstruction and E-mail Writing from a subset of candidates were presented to twenty-one educated native English speakers to be judged for content accuracy and vocabulary usage. Selected item responses to Story Retellings from a subset of candidates were presented to nine educated native English speakers to be judged for content accuracy and vocabulary usage to make the Story Retellings task automatically scorable. Before the raters began rating responses, all were trained to evaluate responses according to analytical and holistic rating criteria. All raters held a master's degree in either applied linguistics or TESOL.

The raters logged into a web-based rating system and evaluated the written responses to Passage Reconstruction and E-mail Writing items for such traits as vocabulary, grammar, organization, and voice and tone. They also evaluated transcriptions of Story Retellings responses, one at a time, for content and vocabulary. The raters' judgments were based on transcriptions instead of recorded spoken responses in order to minimize confounding effects - that is, to ensure that pronunciation or fluency qualities would not affect the evaluation of content and vocabulary. Rating stopped when each item had been judged by three raters. For pronunciation and fluency scoring, the models developed for the Versant English Test were used because those pronunciation and fluency models were trained on a much larger sample of English learners and have proven to be very robust and content independent. Both tests are designed to measure facility in spoken English. Empirical evidence has demonstrated that the Versant English Test is a valid tool to assess spoken English.

6.3 Machine Scoring

Automated scoring methods are used for both the spoken and written constructed responses in VPET. A subset of the responses collected on each item during the field test were subjected to human ratings on various aspects of language skills (traits). These human ratings were then used to train an artificial

intelligence engine. The end result is a set of models able to produce ratings that are predictive of those that expert human raters would give. The model can then be used to score new responses on the same prompts in operational testing, enabling time, effort, and cost savings because human ratings are no longer needed. For an overview of automated scoring technology, see Bernstein, Van Moere, and Chen (2010) for automated scoring for spoken responses, and Foltz, Streeter, Lochbaum, and Landauer (2013) for written response scoring.

7. Validation

The validity of the VPET test scores comes from five different pieces of evidence: (1) the relationship between VPET and PTE Professional scores, (2) standard error of measurement, (3) dimensionality of VPET overall and skill scores, (4) accuracy of the machine score, and (5) the relationship between VPET and VEPT scores. Each piece of evidence is reported separately in the subsequent sections.

7.1 VPET and PTE Professional Scores

VPET has been designed around customers' feedback we collected from PTE Professional and it is intended as an assessment product more closely aligned to English needs typical of an international workplace. PTE Professional was a very reliable, valid assessment tool that measured the ability to understand and use English in the professional context, but concerns had been expressed in terms of test duration and number of items/item types. To address the problem, the test, which originally took about 120 minutes, was shortened and divided into two 60-minute tests of different levels. Each test was shortened by reducing the number of items in some item types (Sentence Completion, Dictation, Repeat, Speaking Situations) and also by completely removing five items types (Picture Description, Summary Writing, Conversations, Passage Reading, Sentence Builds). Therefore, the first question raised in terms of validating the VPET score was whether the VPET scores are strongly related to PTE Professional scores.

In order to answer this question, instead of recruiting a group of participants and having them take both tests, 1,530 PTE Professional tests were selected that had been taken by real test-takers in the past, and then they were re-graded as if they had only taken the reduced form of the test. As a result, it looked as if 1,530 test-takers had taken both tests. These 1,530 tests had been selected to ensure that they included a range of test scores from below-A1 to C2 on the CEFR level. Also, the test-takers represent a range of age and first language backgrounds as well as a representation of both genders. Table 4 summarizes the demographic and proficiency information of the candidates.

Table 4. Description of participants (n = 1530)

Number of Participants	1530 (including 9 native speakers)
Gender	Female: n = 470 Male: n = 889 Unreported: n = 171
Age	Range: 17 to 57 Average: 32 Unreported: 455
First Language	Chinese, English, French, German, Hindi, Japanese, Khasi, Korean, Mongolian, Polish, Portuguese, Punjabi, Russian, Spanish
Proficiency level on CEFR based on PTE Professional score	Below-A1: n = 4 A1: n = 38 A2: n = 322 B1: n = 503 B2: n = 516 C1: n = 125 C2: n = 22

The VPET and PTE Professional scores achieved by the test-takers turned out to be strongly related as shown in the table below.

Table 5. Correlations between VPET and PTE Professional scores (n = 1530)

	Correlation
Overall	.98
Speaking	.92
Listening	.91
Reading	.96
Writing	.98

The correlations between the two test scores are as high as 0.98 in terms of overall score and above .95 in terms of reading and writing scores. The correlation of speaking scores is slightly lower, likely because Passage Reading and Sentence Builds, item types that contributed to the speaking score of PTE Professional, have been completely removed in VPET, and the number of items in Repeat and Speaking Situations has decreased from 14 to 10 and 3 to 2, respectively. Similarly, the lower correlation of listening scores could be attributed to the complete removal of Conversations and the reduced number of items in Dictation (from 14 to 8), the items types that contributed to the listening score of PTE Professional. However, all the correlations are quite high and support the proposition that despite the substantial changes made to the PTE Professional test blueprint, VPET test-takers receive similar or identical scores to what they received on PTE Professional and therefore the new test is essentially measuring the same constructs as the old. This in turn means that conclusions about alignment between scores on PTE Professional and the CEFR, for example, also hold for the new VPET scores.

7.2 Standard Error of Measurement

The standard error of measurement (SEM) provides an estimate of the amount of error due to unreliability in an individual's observed test score and "shows how far it is worth taking the reported score at face value" (Luoma, 2004). If a candidate were to take the same test repeatedly (with no new learning taking place between testings), the standard deviation of his/her repeated test scores is denoted as the SEM. Because no test is perfect, a human variability in performance from moment to moment is inevitable, a test-taker is unlikely to obtain exactly the same score on equivalent (or even identical tests) every time. The SEM of a test's scores is an estimator of this inevitable variability. The SEM of the VPET Overall score has been calculated at 3.1 GSE points. In other words, if a candidate received an Overall score of 50 on VPET, we are 96% confident that this person's "true" overall score falls between 43.8 and 56.2 (3.1 points, doubled, and both added and subtracted from the actual observed score). This low SEM is a result of good alignment among items, consistently accurate scoring, and test reliability.

7.3 Dimensionality: Correlations among Skill Scores

Each skill score on a test ideally provides unique information about a specific dimension of candidate's abilities. For language tests, the expectation is that there will be a certain level of covariance between skill scores given the nature of language learning. When language learning takes place, candidates' skills tend to improve across multiple dimensions simultaneously, but not necessarily in lockstep. For example, someone whose learning was mainly classroom- and textbook-based might spend several weeks on a homestay program in a place where the target language is spoken. Most learners in this situation will show a pattern of a boost in oral skills as a result of this experience that is unlikely to be matched in their written skills, because the experience of suddenly living in the target language environment tends to favor interpersonal communication, pragmatics, and psycholinguistic processing. On the other hand, if all the skill scores were to correlate perfectly with one another, then the skill scores might not be measuring the different aspects of facility with the language originally intended to be the targets of measurement. Table 6 presents the correlations among the VPET skill scores and the Overall score.

Table 6. Inter-correlation between skill scores on the VPET (n = 1530)

	Speaking	Listening	Reading	Writing	Overall
Speaking	-	.70	.65	.73	.86
Listening		-	.71	.76	.92
Reading			-	.80	.87
Writing				-	.91

As expected, skill scores correlate with each other moderately highly by virtue of presumed general covariance within the candidate population between different component elements of language skills. The correlations between the skill scores are, however, significantly below unity, indicating that the different scores measure different aspects of the test construct, using different measurement methods, and different sets of responses. This pattern of results suggests convergent and divergent validity of measurement in VPET scores.

7.4 Machine Accuracy

As mentioned earlier, VPET was created with some of the items originally developed and validated for PTE Professional. Because the accuracy of the machine scoring had been already validated for all those items, a separate validation effort was not made for the VPET with regard to the automated speaking/writing scoring (because it remained unchanged). As a reference, presented below is the summary of the findings on the correspondence between human and machine scores for PTE Professional.

Table 7 shows Pearson Product-Moment correlations between human and machine scores achieved by 124 candidates.

Table 7. Correlation Coefficients between Human and Machine Scoring (n = 124)

Written Proficiency		Oral Proficiency	
Score	Correlation	Score	Correlation
Writing Profile	.98	Speaking Profile	.95
Grammar	.99	Sentence Mastery	.93
Word Choice	.98	Vocabulary	.95
Organization	.90	Fluency	.85
Voice & Tone	.91	Pronunciation	.84
Reading	.96	Listening	.96

For PTE Professional, the scores of four subskills (grammar, word choice, organization, voice & tone) underlying writing ability were reported separately. These writing-related scores were combined with reading scores to calculate and report a writing profile score, which represented the level of written proficiency. Likewise, the scores of four subskills (sentence mastery, vocabulary, fluency, pronunciation) underlying speaking ability were presented separately on the score report. The speaking related scores were combined with listening score to derive a speaking profile score, which represented the level of oral proficiency. Therefore, correlations are shown here for each score that was shown on the PTE Professional score report, some of which do not appear on the VPET score report.

The human scores were calculated from a single human judgment, which means that the correlation coefficients are low-end estimates, because higher coefficients can be obtained with multiple human ratings which wash out inter-rater variability. Table 7 demonstrates that the Writing Profile scores of PTE Professional produced automatically by machine yielded scores that closely corresponded with human ratings. Among the subscores, the human-machine relation is closer for the linguistic (Grammar and Word Choice) and content (Reading) aspects of written language than for the rhetorical aspect (Organization and Voice & Tone), but the relation is close for all five written proficiency subscores.

The table also suggests that the Speaking Profile scores of PTE Professional by machine yielded scores that closely corresponded with human ratings. Among the subscores, the human-machine relation is closer for Listening and the content aspects of spoken language (Sentence Mastery and Vocabulary) than for the manner of speaking subscores (Fluency and Pronunciation), but the relation is close for all five oral proficiency subscores. In summary, these patterns of human-machine scoring

correspondence demonstrate that at the profile score level, machine-generated scores are virtually indistinguishable from scoring done by careful human transcriptions and human judgments.

7.5 VPET and VEPT

Another validation effort has been made to understand how VPET scores relate to other measures of English proficiency. Because VPET measures all four skills (Speaking, Listening, Reading, Writing) and the scores represent a candidate's English proficiency in everyday workplace contexts, VEPT (Versant English Placement Test) was used for a concurrent validation experiment. VEPT is also a four-skill test, and it measures how well a person can understand and use English on everyday topics. VEPT and VPET share some item types including Repeat, Sentence Completion, Dictation, and Passage Reconstruction, although the contents or topics of the items differ to some extent.

7.5.1 Method

The study was conducted between August 2020 and November 2020. Originally, 81 participants were recruited to take both VPET and VEPT. However, only 57 participants' test data were included in the analysis because the data from 24 participants were deemed unsuitable, as they did not complete the test(s) or skipped a large number of items. Of the 57 participants, 25 were male and 32 were female with a mean age of 27. They have diverse first language backgrounds (e.g., Bahasa, Spanish, Hindi, German, Chinese, Slovenian, Portuguese).

7.5.2 Results

Table 8. Correlations between VPET and VEPT scores (n = 57)

	Correlation
Listening	.80
Reading	.75
Speaking	.77
Writing	.80
Overall	.88

As would be expected when test-takers are given different tests of somewhat overlapping constructs, the pattern of correlations demonstrates a moderately but not perfectly high correspondence between scores on the two tests. Correlations below .50 or so would indicate a substantial divergence between scores on the two tests, but correlations at .75 or .80 as shown are precisely what might be expected for tests which measure similar but different constructs using similar but different item types.

8. Conclusion

This report has provided validity evidence to assist test score users to make an informed interpretive judgment as to whether or not VPET scores would be valid for their purposes. The test development process is documented and adheres to sound theoretical principles and test development ethics from the field of applied linguistics and language testing. In particular, the items were written to specifications

and subject to a rigorous procedure of qualitative review and psychometric analysis before being deployed to the item pool; the content was selected from both pedagogic and authentic material; the test has a well-defined construct that is represented in the cognitive demands of the tasks; the scoring weights and scoring logic are explained; the items were widely field tested and analyzed on a representative sample of candidates and psychometric properties of items are demonstrated; and further, empirical evidence is provided which verifies that VPET scores are structurally reliable indications of candidate ability in spoken and written English and are suitable for decision-making.

9. About the Company

Pearson: Pearson and Ordinate Corporation, the creator of the Versant tests, were combined in January, 2008. The Versant tests are the first to leverage a completely automated method for assessing spoken and written language.

Versant Testing Technology: The Versant automated testing system was developed to apply advanced speech recognition techniques and data collection to the evaluation of language skills. The system includes automatic mobile phone and computer reply procedures, dedicated speech recognizers, speech analyzers, databanks for digital storage of speech samples, and score report generators linked to the Internet. VPET is the result of years of research in speech recognition, statistical modeling, linguistics, and testing theory. The Versant patented technologies are applied to its own language tests such as the Versant series and also to customized tests. Sample projects include assessment of spoken English, children's reading assessment, adult literacy assessment, and collections and human rating of spoken language samples.

Pearson's Policy: Pearson is committed to the best practices in the development, use, and administration of language tests. Each Pearson employee strives to achieve the highest standards in test publishing and test practice. As applicable, Pearson follows the guidelines propounded in the Standards for Educational and Psychological Testing, and the Code of Professional Responsibilities in Educational Measurement. A copy of the Standards for Educational and Psychological Testing is available to every employee for reference.

Research at Pearson: In close cooperation with international experts, Pearson conducts ongoing research aimed at gathering substantial evidence for the validity, reliability, and practicality of its current products and investigating new applications for Versant technology. Research results are published in international journals and made available through the Versant website (www.VersantTests.com).

10. References

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

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377.

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

Bull, M & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In R.H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing*. Canberra, Australia: Australian Speech Science and Technology Association.

Carver, R. (1991). Using Letter-naming speed to diagnose reading disability. *Remedial and Special Education, 12*(5), 33-43.

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

Council of Europe (2009). *Manual for relating language examinations to the common European Framework of Reference.* Cambridge, UK: Cambridge University Press.

Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy – Mental imagery, philosophical issues about*. London: Nature Publishing Group, 858-864.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). New York: Routledge, 68-88.

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

Gough, P. B., Ehri, L. C., & Treiman, R. (1992). *Reading acquisition.* Hillsdale, NJ: Erlbaum.

Grabe, W., and Kaplan, R.C. (1996). *Theory and practice of writing.* New York: Longman.

Jescheniak, J.D., Hahne, A. & Schriefers, H.J. (2003). Information flow in the mental lexicon during speech planning: Evidence from event-related brain potentials. *Cognitive Brain Research, 15*(3), 261-276.

Landauer, T.K., Foltz, P.W. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. *PNAS, 98*(23)*,* 13464-13471.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2,* 217-228.

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, Inc. (2017). Pearson Test of English – Professional: Test description and validation summary.

Perry, J. (2001). *Reference and reflexivity*. Stanford, CA: CSLI Publications.

Segalowitz, N., Poulsen, C., & Komoda, M. (1991). Lower level components or reading skill in higher level bilinguals: Implications for reading instruction. In J.H. Hulstijn and J.F. Matter (eds.), *Reading in two languages*, AILA Review, Vol. 8,. Amsterdam: Free University Press, 15-30.

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.

Van Turennout, M., Hagoort, P. & Brown, C. M. (1998). *Brain Activity During Speaking: From Syntax to Phonology in 40 Milliseconds. Science, 280,* 572-574.

About Us

We are Pearson English, part of the world's learning company, with expertise in educational courseware and assessment, and a range of teaching and learning services powered by technology.

With 30,000 employees in more than 70 countries, our products are used by millions of professionals, teachers and learners around the world every day. Whether you're a learner seeking swift progress towards new horizons, a teacher who's inspiring achievement in the classroom, an institution looking for measurable improvement, or a professional striving to make data-backed decisions and upskill and reskill their talent for the future, the world of language learning is evolving.

Our mission is to help people make progress in their lives through learning – because we believe that learning opens up opportunities, creating fulfilling careers and better lives.

To try a sample test or get more information, visit us online at:

www.VersantTests.com

Version 0822