Versant™ English Placement Test

Test Description and Validation Summary

Table of Contents

1. Introduction

The Versant™ English Placement Test (VEPT), powered by Versant technology, is an assessment instrument designed to measure how well a person can understand and use English on everyday topics. The VEPT is intended for adults and students over the age of 16, and takes approximately 50 minutes to complete. Because the VEPT is delivered automatically, the test can be taken at any time, from any location via computer. A human examiner is not required. The computerized scoring allows for immediate, objective, and reliable results that correspond well with traditional measures of spoken and written English performance.

The VEPT measures *facility* in spoken and written English which is how well a person can understand spoken and written English and respond appropriately in speaking and writing on everyday topics, at an appropriate pace in intelligible English. VEPT scores provide reliable information that can be used for such decisions as placement, exit from intervention, and progress monitoring by academic and government institutions as well as commercial and business organizations.

2. Test Description

2.1 Test Design

The VEPT has eight automatically scored tasks: Read Aloud, Repeats, Sentence Builds, Conversations, Sentence Completion, Dictation, Passage Reconstruction, and Summary & Opinion. These tasks provide multiple, fully 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. Because more than one task contributes to each skill score, the use of multiple tasks strengthens score reliability. The VEPT also includes a Typing task that is not scored but provides information of the typing speed and accuracy.

The VEPT provides numeric scores and performance levels that describe the candidate's facility in spoken and written English. The VEPT score report is made up of an Overall score and four skill scores: Speaking, Listening, Reading, and Writing. The Overall score is an average of the four skill scores. Together, these scores describe the candidate's facility in spoken and written English. As supplemental information, Typing Speed and Typing Accuracy are also reported on the score report.

2.2 Test Administration

The VEPT is administered via Versant for Web (VfW), a browser-based system. It is available in both an on-line and off-line mode. The VEPT 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 necessary to verify the candidate's identity and/or to ensure that the test is taken under exam conditions.[1]

Administration of a VEPT generally takes about 50 minutes. It is best practice for the administrator to give a test paper to the candidate at least five minutes before starting the test (see Appendix for a sample test paper). The test paper contains instructions for each of the nine tasks. The candidate then has the opportunity to read the test paper and ask questions before the test begins. At this time, the administrator can answer any procedural or content-related questions that the candidate may have.

The candidate must use a microphone headset to take the VEPT in order to guarantee a consistent sound quality of both test content and responses. When the test is launched, the candidate is prompted to enter either a unique Test Identification Number (TIN) or his/her email address which is associated with a unique TIN. The TIN is provided on the test paper and on ScoreKeeper, a secure Pearson website. The candidate is prompted to adjust the volume to an appropriate level and to test the microphone before beginning the test.

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 or typing on the computer keyboard.

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. A timer is shown in the upper right-hand corner of the computer screen. 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.

2.3 Test Format

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

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.

[1] For more secure testing, the VEPT can also be administered via Pearson's Computer Delivered Test (CDT) software, which "locks down" the computer to prevent web browsing, consulting files on the local hard drive, copying or pasting, etc.

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: Read Aloud

In the Read Aloud task, candidates are asked to read two short passages out loud, one at a time. Candidates are given 30 seconds to read each passage. The texts are displayed on the computer screen.

The passages are expository texts that deal with general everyday topics. All passages are relatively simple in structure and vocabulary and range in length from 60 to 70 words. The SMOG (Simple Measure of Gobbledygook) Readability Index was used to identify and refine the readability level of each passage (McLaughlin, 1969). SMOG estimates the number of years of education needed to comprehend a passage, and the algorithm factors in the number of polysyllabic words across sentence samples. All Read Aloud passages have a readability score between 6 and 8 so they can be read easily and fluently by most educated English speakers.

For candidates with little facility in spoken English but with some reading skills, this task provides samples of 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 screen and provide information about how well candidates 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 candidates, reading aloud presents a familiar task and is a comfortable introduction to the interactive mode of the test as a whole.

Example:

> Australia is a very large country. It is the sixth largest country in the world. It is also a continent and is sometimes called the 'island continent'. It is surrounded by two oceans. Most of Australia is a desert so it is very flat and dry, but it also has rain forests and mountains. It is home to many different kinds of animals.

Part B: Repeats

In this task, candidates are asked to repeat verbatim sentences spoken to them through their headphones. The sentences are presented 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.

Examples:

1. He's a great teacher.
2. It's not too late to change your mind.
3. People know how easy it is to get lost in thought.

To repeat a sentence longer than about seven syllables, a person must recognize and parse a continuous stream of speech into words (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 less familiar with English sentence structure.

Because Repeat items require candidates to recognize what they heard, then represent what was said in linguistic units, they 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 C: Sentence Builds

For the Sentence Builds task, candidates hear three short phrases and are asked to rearrange them to make a sentence. The phrases are presented in a scrambled order and the candidate mentally rearranges them, then constructs and says a sentence made up of the three phrases.

Examples:

1. my boss / to London / moved
2. of your family / any pictures / do you have
3. to their leader / listened carefully / the young men

To correctly complete this task, a candidate must understand the meaning of the individual 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 phrase that a person can hold in verbal working memory. This is important to measure because it reflects the candidate's ability to access and retrieve lexical items and to build phrases and clause structures automatically. The more automatic these processes are, the greater the candidate's facility in spoken English. This skill is demonstrably distinct from memory span (as further discussed in Section 2.5.2).

The Sentence Builds task involves constructing and articulating sentences. As such, it is a measure of candidates' mastery of sentences, in addition to their pronunciation and fluency.

Part D: Conversations

In the Conversations task, candidates listen to a conversation between two speakers, which typically consists of three speaking turns. Immediately after the conversation, candidates are asked a comprehension question, then answer the question by saying a word or short phrase.

Example:

> Speaker 1: Congratulations on graduating!
> Speaker 2: Thanks! It was a lot of work.
> Speaker 1: I know. You deserve a party.
>
> Question: Why does the man deserve a party?

This task measures candidates' listening comprehension ability. Conversations are recorded at a conversational pace covering a range of topics. The task requires candidates 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 questions.

Part E: Typing

The VEPT includes a typing speed and accuracy task which is not included in the actual test scores. In this task, candidates see a passage on the computer screen and have 60 seconds to type the passage exactly as it appears. 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 level of 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.

Example:

> Many people do not like public speaking. They are afraid to speak in front of a large group of people. There are many ways to get better at public speaking. First, it is good to know the room. You should know where to stand and where to set up your computer. Second, it is important to know the audience. If you greet some people as they arrive, you will feel more comfortable because you will be familiar with them. Lastly, you need to be prepared. You should practice your speech as much as you can and revise it if necessary.

This task has several functions. First, it allows candidates to familiarize themselves with the keyboard prior to the actual writing tasks. Second, it measures the candidate's typing speed and accuracy. The VEPT assumes a basic competence in typing for every candidate. Because it is important to disambiguate candidates' typing skills from their written English proficiency, it is recommended that test score users review each candidate's typing score. If typing speed is below 12 words per minute, and/or accuracy is below 90%, then it is likely that this candidate's written English proficiency was not properly measured

due to inadequate typing skills. The test administrator should take this into account when interpreting test scores.

Part F: Sentence Completion

In the Sentence Completion task, candidates read a sentence that has a word missing, then they supply an appropriate word to complete the sentence. Candidates are given 25 seconds for each item. During this time, candidates must read and understand the sentence, think of an appropriate word, and type the word in the text box provided to complete the sentence. Sentences range in length from 5 to 25 words. In many Sentence Completion items, there is more than one possible correct answer. Across all items in this task, candidates encounter sentences with words missing from various parts of speech (i.e., nouns, verbs, adjectives, adverbs) and from different positions in sentences: sentence-initial, sentence-medial, sentence-final.

Examples:

1. Her favorite hobby is _____. She has so many books.
2. He arrives _____ and is often the first one here.
3. I asked a coworker to take over my _____ because I wasn't feeling well.

It is sometimes thought that fill-in-the-gap tasks (in some cases also called cloze tasks) are more authentic when longer passages or paragraphs are presented to the candidate, as 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.

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."

The Sentence Completion task draws on interpretation, inference, lexical selection and morphological encoding, and as such reflects the candidate's mastery of vocabulary in use.

Part G: 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 3 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. I'll see you on Thursday.
2. How long can I keep this book?
3. She apologized to all her friends several times.

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), which refers to the range of contextually-influenced choices made by language users. Proficient listeners tend to understand and remember the content of a message, but not the exact words used; they retain the meaning of the message rather than the words used to 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 from their memory of its contents. 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).

Dictation is a good test task of comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for candidates less 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. Candidates are asked to read a text, put it aside, and then write out what they 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 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, and results.

In order to perform this task, the candidate must read the passage presented, understand the narrative, and hold the concepts and details in memory long enough to reconstruct the passage accurately in writing. Individual candidates may naturally employ various strategies when performing this task. Reconstruction may be more or less 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 responding to requests in writing, replying to emails, recording events or decisions, or summarizing texts.

Example:

> Robert went to a nice restaurant for dinner. When the waiter brought the bill, Robert reached for his wallet, but it wasn't in his pocket. He remembered having his wallet when he came into the restaurant. The waiter looked around the floor near his table. He found the wallet under the table.

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 first language, 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 I: Summary & Opinion

In the Summary & Opinion task, candidates are presented with a reading passage. They are given 18 minutes to read it, 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. 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 is used to identify and adjust the readability level of each passage. All passages have a readability score around 10, and are therefore easily understandable by most educated English speakers.

Example:

Some children grow up in a big city while other children grow up in the countryside. Childhood experiences can be very different depending on where a person is raised. Although the countryside can be more peaceful than a big city, it is better for children to grow up in a big city.

Children who grow up in a big city have more opportunities. If a child wants to sing, dance, or play a musical instrument, he or she can easily find different teachers or clubs. A child who is interested in sports has a lot of sports to choose from. In addition, most big cities have excellent museums, zoos, art galleries, and libraries. Therefore, children can spend their evenings, weekends, and summers learning about many different subjects. By experiencing a wide range of activities, children will be able to find out what they like and maybe find a special interest.

Children can develop a world view in a big city. In big cities, there are people from many different backgrounds. It is good for young people to meet people from different cultures. It prepares them for the real world. They can learn ideas or opinions that are different from the ones they are used to. By meeting people from all over the world, a big city helps children to understand how different people communicate. When faced with a problem, a big city child is more likely to consider many different solutions.

Some people do not like big city life, but it has more opportunities and more culture than life in the countryside. Living in a big city is a great way to prepare children for the real world.

Write a short summary of the author's opinion in 25-50 words. Do not copy.

Write your opinion. Do you agree or disagree with the author? Why? Write at least 50 words. Try to use your own ideas.

In the Summary response, candidates 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 candidate must read the passage, understand the concepts and details, and evaluate the content by identifying the most important points. Candidates must construct an informative and succinct response with appropriate spelling, punctuation, capitalization, syntax, and grammar. Responses are scored on the quality of the summary and on adherence to English writing conventions.

In the Opinion response, candidates are expected to provide their own opinion on the topic presented and to provide clear and appropriate supporting ideas and/or examples. Candidates must construct an informative response with appropriate spelling, punctuation, capitalization, syntax, and grammar. Responses are scored on the quality of the opinion and on adherence to English writing conventions.

The Summary & Opinion task draws on reading comprehension, interpretation, inference, summarization, syntax, and writing mechanics, and as such reflects the candidate's mastery of reading and writing.

2.4 Number of Items

In the administration of the VEPT, the testing system presents approximately 81 items in nine separate sections to each candidate. 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 task

Task	Approximate Number of Items
A. Read Aloud	2
B. Repeats	16
C. Sentence Builds	10
D. Conversations	12
E. Typing	1
F. Sentence Completion	20
G. Dictation	16
H. Passage Reconstruction	3
I. Summary and Opinion	1
Total	**81**

2.5 Test Construct

2.5.1 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). The VEPT is designed to measure a candidate's facility in spoken and written English—that is, the ability to understand spoken and written English and respond appropriately in speaking and writing on everyday topics, at an appropriate pace 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. Repeats, Sentence Builds, Conversations, and Dictation expose candidates to spoken English, and Read Aloud, Sentence Completion, Passage Reconstruction, and Summary & Opinion present written English that candidates must read and comprehend within given time limits.

Repeats, Sentence Builds, Conversations, 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. In particular, Buck (2001) asserts that dictation is not so much an

assessment of listening skills, as it is sometimes perceived, but is rather an assessment of general language ability, requiring both receptive and productive knowledge. This is because it involves both comprehension and (re)production of accurate language.

Sentence Completion, Passage Reconstruction and Summary & Opinion 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). Alderson (2000) suggests that these comprehension skills involve vocabulary, discourse and syntactic knowledge, and are therefore general linguistic skills which may pertain to listening and writing as much as they do to reading.

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 VEPT 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).

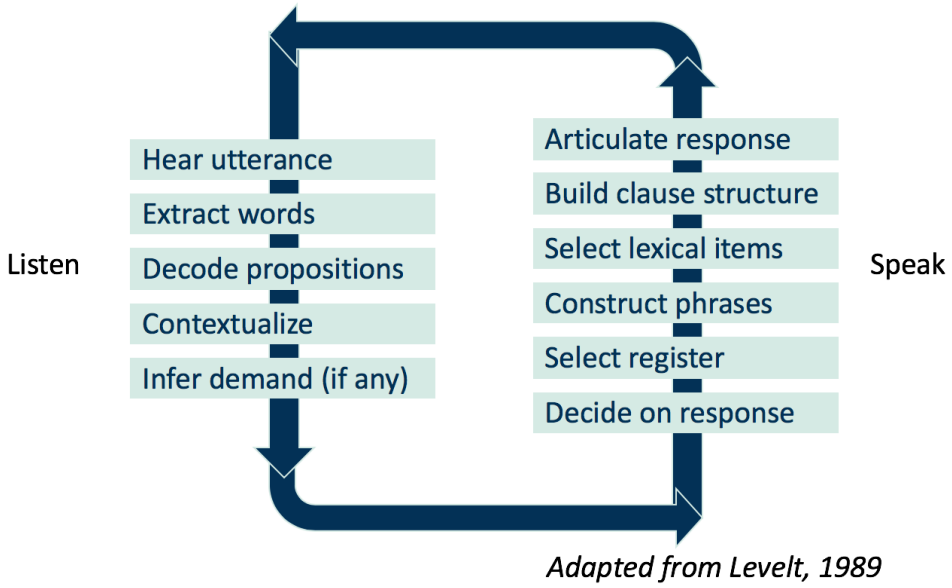


Figure 1. Conversational processing components in listening and speaking

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 Summary & Opinion, 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 for discriminating candidates with different levels of *automaticity*.

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 VEPT probes the degree of automaticity in language performance.

By utilizing integrated tasks, the VEPT 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 Repeats, Sentence Builds, Sentence Completion and Passage Reconstruction tasks of the VEPT as candidates are asked to anticipate missing words and reconstruct texts.

2.5.2 The Role of Memory

Some measures of automaticity can be misconstrued as memory tests. Because some VEPT tasks involve repeating long sentences, holding phrases in memory in order to assemble them into reasonable 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, Sentence Build, 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 spans.

2.5.3 The Role of Context

The VEPT probes the psycholinguistic elements of spoken and written language performance rather than the social, rhetorical, and cognitive elements of communication. All items, with the exception of Summary & Opinion, present context-independent material in English. Context-independent material is used in the majority of 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 speaking and writing the language because less time is spent presenting contexts that situate a language sample or set up a task demand.

In summary, there are many processing elements required to participate in spoken and written exchanges of communication: a person has to recognize spoken or written words, understand the message, formulate a relevant response, and then produce an appropriate response at an acceptable pace in intelligible English. Accordingly, the constructs that can be observed in the candidate's performances in the VEPT are knowledge of the language, such as grammar and vocabulary, comprehension of the information conveyed through the language, knowledge or spoken production, such as pronunciation and stress, and knowledge of writing conventions, such as organization and spelling. Underlying these observable performances are psycholinguistic skills such as *automaticity* and *anticipation*. As candidates operate with spoken and written English and select words for constructing sentences, those able to draw on many hours of relevant experience with grammatical sequences of appropriate words will perform at the most efficient speeds.

3. Content Design and Development

3.1 Vocabulary Selection

The vocabulary used in all spoken test items and responses is restricted to forms of the 5,000 most frequently used words in the Switchboard Corpus (Godfrey & Holliman, 1997), a corpus of three million

words spoken in spontaneous telephone conversations by over 500 speakers of both sexes from every major dialect of American English. In general, the language structures used in the test reflect those most 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.” The vocabulary used in all written test items and responses is restricted to forms of the 1,600 most frequently used words in the Longman Corpus Network, a database of 430 million words of spoken and written English, collected from both British and American English sources.

3.2 Item Development

VEPT items were drafted by native English-speaking item developers from different regions in the U.S. In general, the language structures used in the test reflect those that are common in everyday spoken and written English. The items were designed to be independent of social nuance and complex cognitive functions.

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, represented common English usage in different English-speaking regions, and contained appropriate content. Then, draft items were sent to external linguists for expert review to ensure 1) compliance with the vocabulary specification, and 2) conformity with current English usage in different countries. Reviewers checked that items would be appropriate for candidates trained to standards other than American English.

All items were checked for compliance with the vocabulary specification. Most vocabulary items that were not present in the lexicon were changed to other lexical stems that were in the consolidated word list. Some off-list words were kept and added to a supplementary vocabulary list, as deemed necessary and appropriate. 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 90% of a reference sample of educated native speakers of English.

3.3 Item Prompt Recording

3.3.1 Voice Distribution

Forty-three native speakers (22 female and 21 male) representing various speaking styles and regions such as the U.S., U.K., and Australia, were selected for recording the spoken prompt materials. Some items were also recorded by non-native speakers of English. Care was taken to ensure that the non-native speakers were at advanced levels in terms of their speaking ability, and that their pronunciation was clear and intelligible. These speakers’ country of origin included China, Costa Rica, India, Israel, Italy, and Korea.

Recordings were made in a professional recording studio. 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 prompt voices.

3.3.2 Recording Review

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

4. Score Reporting

4.1 Scores and Weights

The VEPT score report is comprised of an Overall score and four skill scores (Speaking, Listening, Reading, and Writing).

Overall: 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 an appropriate pace and in intelligible English. Scores are based on a weighted combination of the four skill scores. Scores are reported in the range from 10 to 90 on Pearson's Global Scale of English (GSE). The corresponding Common European Framework of Reference for Languages (CEFR) level is also displayed.

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.

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

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

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

In the VEPT 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.

Each incoming spoken response from a VEPT 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 Read Aloud, Repeats, Sentence Builds, and Conversations 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 VEPT 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 Summary and Opinion 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 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 them and posts scores at www.VersantTest.com. Test administrators and score users can view and print out the test results from ScoreKeeper.

Score users may be educational and government institutions as well as commercial and business organizations. Pearson endorses the use of Versant test scores 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

VEPT themselves under secure conditions, or by having trusted third parties administer the test. In several countries, education and commercial institutions provide such services.

VEPT scores can be used to assess how well and efficiently a candidate can process and produce spoken and written English on everyday topics. VEPT scores can be used to evaluate the level of spoken and written English skills of individuals entering into, progressing through, or leaving English language courses. The VEPT score scale covers a wide range of abilities in spoken and written English communication; therefore, it is effective for placement purposes, progress monitoring, and exit testing.

It is up to score users to decide what VEPT score can be regarded as a minimum requirement in their context (a "cut score"). Score users may wish to base their selection of an appropriate criterion score on their own localized research. Pearson can provide assistance in helping organizations to arrive at data-based criterion scores.

5. Field Testing

Both native speakers of English and English language learners were recruited as participants to take a field test version of VEPT. The purposes of this field testing were 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.

5.1 Native Speakers

A total of 161 educated adult native English speakers were recruited. Most were from the U.S. with a few from the U.K., Canada, Australia, and New Zealand. The male to female ratio was roughly equal.

While VEPT is 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

A total of 1,194 English language learner candidates were recruited from various countries, including Argentina, China, India, Indonesia, Italy, Japan, Korean, Philippines, Russia, Singapore, and Turkey. A total of 45 different languages were reported. The male to female ratio was roughly equal.

6. Data Resources for Scoring Development

As a result of the field test, more than 147,801 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 responses were transcribed at least once, and the majority of the English learner responses were transcribed two or more times to ensure 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 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 transcriptions and learning a specific set of rules and annotation symbols. Subsequently, they completed a series of training sets. Only the transcribers who met 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.

There were two steps to the transcription task. The first step involved producing transcriptions to develop automated scoring systems. For this step, a total of 39,416 transcriptions were produced (12,119 transcriptions for L1 English speaker responses and 27,297 transcriptions for learner responses.) The second step was to produce transcriptions to conduct a series of validation analyses. An additional 20,511 transcriptions were produced for the validation analysis purpose.

6.2 Human Rating

Field test responses were scored 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. Responses to Passage Reconstruction and Summary & Opinion were presented to thirty-five educated native English speakers. For the Passage Reconstruction responses, the raters judged for Narrative

Clarity and Accuracy. For the Summary & Opinion responses, the raters judged for Summary Writing, Opinion Writing, and Writing Conventions. Before the raters began rating responses, they all underwent rigorous training, which included reviewing rating materials and completing several training sets. Raters who did not meet the standards were not involved in the actual scoring process. During the actual rating process, the ratings were reviewed by the test development team and the raters were given feedback throughout the process.

The actual rating process was carried out using the same online interface used by the transcribers. Responses were presented from a stack semi-randomly, so that a single candidate's set of responses would be spread among many different raters, similar to the transcription process. For Summary Writing, Opinion Writing, and Writing Conventions, two scores were collected by two independent raters, and with the Narrative Clarity and Accuracy rating, each response was scored by three independent raters. For pronunciation and fluency scoring, the models developed for the Versant English Test[2] were used because those models were trained on a very large sample of data, 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.

As in the transcription task, the rating task also consisted of two steps. The first part was to collect expert judgments to develop automated scoring systems and the second step was to conduct a series of validation analyses. The experts produced a total of 82,131 ratings for the development of automated scoring systems and a total of 27,275 ratings for the validation analysis purpose.

6.3 Machine Scoring

Automated scoring methods are used for both the spoken and written constructed responses in VEPT. 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) as described above. 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 need. 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

Trained scoring models are generally successful at reproducing the human ratings they have been trained on. However, in operational testing, the automated scoring system needs to deal with new candidates and responses that the machine has never encountered before. It is, therefore, critical to demonstrate how closely machine scores correspond to human scores on a new set of data that is

[2] For more information about the Versant English Test, refer to the report *Versant English Test: Test Description & Validation Summary*.

separate from the data used to develop the automated scoring system. If machine scores show close correspondences with human-based scores, it will serve as a piece of validity evidence for the accuracy of VEPT's machine scoring. All scores, statistics, and results in the validation studies below (§7.1-7.3) use the original Versant scale of 20 to 80 rather than the GSE of 10 to 90.

7.1 Validity Study Design

From the large body of spoken and written performance data collection from native speakers of English and learners of English during the field tests, score data were analyzed to provide evidence for the validity of VEPT as a measure of proficiency in spoken and written English. Validation analyses examined several aspects of VEPT scores:

Internal Validity

1. Reliability: the extent to which VEPT is reliable and internally consistent
2. Machine Accuracy: the extent to which the automatically scored VEPT scores are comparable to the scores that human listeners and raters would assign
3. Dimensionality: the extent to which the four different skill scores of the VEPT are sufficiently distinct
4. Differentiation among known populations: the extent to which VEPT Overall scores reflect expected differences between English language learners and native English speakers,

Concurrent Validity

1. Relation to framework with related construct: how VEPT scores correspond to IELTS and TOEFL

7.1.1 Validation Dataset

A total of 214 participants were recruited for a series of validation analyses. These participants were recruited separately from the field test candidates. Care was taken to ensure that the field test dataset and validation dataset did not overlap for independent analyses. This means that the performance samples provided by the validation candidates were excluded from the datasets used for training the automatic speech processing models or for training the scoring models.

Validation participants were recruited from a variety of countries, first language backgrounds, and proficiency levels and were representative of the candidate population using the VEPT. A total of five native English speakers were included in the validation dataset. Table 3 summarizes the demographic information of the validation participants.

Table 3. Description of participants in the validation dataset (*N* = 214)

Number of Participants	214 (including 5 native speakers)
Gender	Female: *n* =107 Male: *n* = 107
Age	Range: 16 to 64 Average: 26 6 unreported
First Languages	Arabic, Bahasa, Burmese, Cantonese, Czech, English, Farsi, Filipino, German, Hindi, Indonesian, Italian, Japanese, Khmer, Swahili, Korean, Malay, Mandarin, Mongolian, Pashto, Portuguese, Punjabi, Russian, Serbian, Tamil, Thai, Turkish, Vietnamese

7.1.2 Descriptive Statistics

Table 4 summarizes the descriptive statistics of the validation dataset test scores. The mean Overall score of the validation dataset is 46.36 with a standard deviation of 13.72 (on a scale of 20-80), indicating that the candidates' scores are spread out across the entire Versant scale.

Table 4. Descriptive statistics for the validation dataset (*N* =214)

Measure	Statistic
Mean	46.47
Median	46.76
Standard Deviation	14.02
Sample Variance	196.54
Kurtosis	-0.78
Skewness	0.00

Figure 2 shows the distribution of the candidate's Overall scores. Candidates' scores are spread out mostly evenly across the entire Versant scale with at least one candidate at every score point.

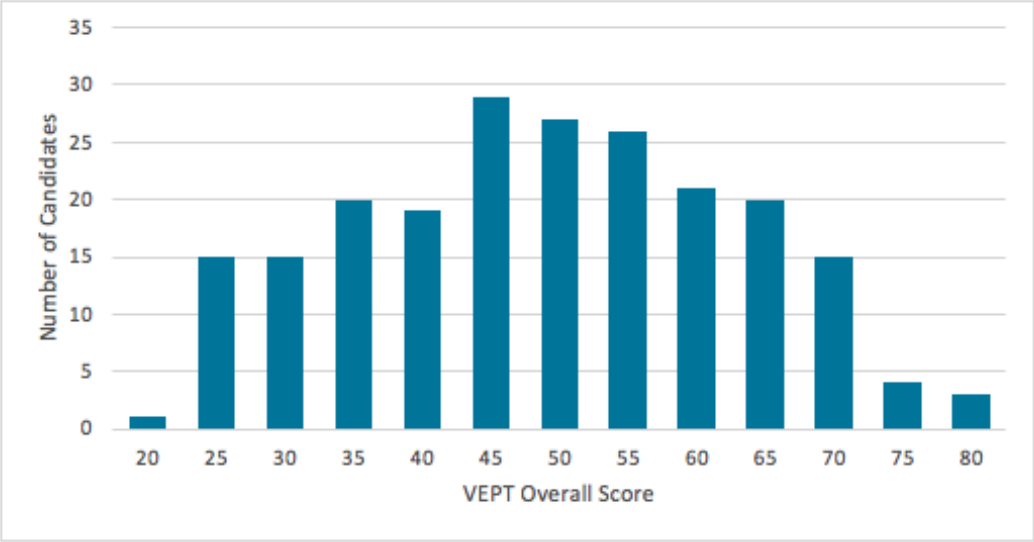


Figure 2. Distribution of VEPT Overall scores of candidates in the validation dataset (*N* = 214).

7.2 Internal Validity

7.2.1 Reliability

Standard Error of Measurement

The standard error of measurement 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, p.183). 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 standard error of measurement. The standard error of measurement of the VEPT Overall score is 2.7. In other words, if a candidate received an Overall score of 50 on VEPT and then took the test again, his or her Overall score is expected to fall between 48.7 and 51.3 on the second test.

Split-Half Reliability

Split-half reliability was calculated for the Overall score and all skill scores with the Spearman-Brown Prophecy Formula used to correct for underestimation. This analysis was conducted on both machine-generated and human-based scores. Table 5 presents the split-half reliability estimates based on the same validation dataset (*N* = 214) scored by careful human rating and transcription in one case, and by automated scoring in the other.

Table 5. Split-half reliability estimates of VEPT machine scores versus human scores (*N* = 214)

Score	Split-half Reliability for Machine Scores	Split-half Reliability for Human Scores
Overall	.99	.99
Speaking	.94	.95
Listening	.95	.97
Reading	.95	.97
Writing	.99	.98

The values in Table 5 suggest that the effect on reliability of using automated scoring technology, as opposed to careful human rating, is very small across all score types. This analysis demonstrates that the VEPT is a highly reliable test and that the reliability of the automated test scores is nearly identical to that of expert human ratings.

Test-Retest Reliability

Score reliability was also estimated by the test-retest method. VEPT is randomized, meaning that a set of test items conforming to the test blueprint is randomly selected from an item pool by the Versant testing system, thereby generating a different "test form" for each candidate on each occasion. If the various VEPT forms do not yield similar test scores, the validity of score interpretations is called into question. Therefore, it is critical to investigate whether candidates receive comparable scores on all VEPT forms.

To examine the reliability, 44 participants were recruited to take two VEPT test forms within a ten-day period. The first test administration is referred to as Test 1 and the second test administration is referred to as Test 2. All Test 1 and Test 2 forms were randomly generated and unique for each test-taker. Both tests were delivered via CDT.

As shown in Table 6, mean test scores on Test 2 were slightly higher than those on Test 1 in all scoring domains, and the slight increase may be attributable to greater familiarity with the test on the second test occasion. However, mean score differences are less than 1.9 points (on Versant's 20 to 80 point scale) and such a deflection is well within the standard error of measurement (2.7 points) of the VEPT. This analysis demonstrates that performance on a VEPT does not differ significantly from one test administration to another when no learning has taken place.

Table 6. Mean Scores on Tests 1 and 2

Score Type	Test 1 Mean	Test 2 Mean	Difference
Overall	49.6	51.2	1.6
Speaking	44.5	45.4	0.9
Listening	52.2	54.1	1.9
Writing	50.5	51.8	1.3
Reading	50.9	51.4	0.5

Test-retest reliability was estimated by calculating Pearson product-moment correlation coefficients for the Overall score and the four subskill scores (Speaking, Listening, Writing, and Reading). Results of the correlation analyses are summarized in Table 7.

Table 7. Pearson correlations for Versant English Placement Test Overall and Subskill Scores (n = 44)

Score Type	r
Overall	.95
Speaking	.91
Listening	.86
Writing	.87
Reading	.79

Test-retest reliability for the Overall scores is .95. The results of this study present convincing evidence that VEPT scores are highly comparable and consistent across multiple test occasions, regardless of the form test-takers received.

7.2.2 Machine Accuracy

Another analysis for internal quality involved comparing scores from the VEPT, which uses automated language processing technologies, with scores derived from human transcriptions and human judgments by expert raters.

Table 8 presents Pearson Product-Moment correlations between machine scores and human scores, when both methods are applied to the validation dataset. Correlations presented in Table 10 suggest that scoring a VEPT by machine yields scores that closely correspond with human ratings.

Table 8. Correlations between human and machine scoring of VEPT responses (N = 214)

Score Type	Correlation
Overall	.98
Speaking	.91
Listening	.98
Reading	.98
Writing	.90

At the Overall score level, VEPT machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and multiple independent human ratings.

7.2.3 Dimensionality: Correlations Among Skill Scores

Each skill score on a test ideally provides unique information about a specific dimension of the candidate's ability. 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, the candidate's skills tend to improve across multiple dimensions. However, if all the skill scores were to correlate perfectly with one another, then the skill scores might not be measuring different aspects of facility with the language. A dataset of 1,000 VEPT tests was randomly selected from tests delivered over a two-month period. A broad range of native languages is represented. Table 9 presents the correlations among the VEPT skill scores and the Overall score for this sample.

Table 9. Inter-correlation between skill scores on the VEPT (n = 1,000)

	Speaking	Listening	Reading	Writing	Overall
Speaking	-	.81	.64	.63	**.86**
Listening		-	.76	.76	**.93**
Reading			-	.77	**.88**
Writing				-	**.89**

As expected, skill scores correlate with each other to some extent by virtue of presumed general covariance within the candidate population between different component elements of spoken language

skills. The correlations between the skill scores are, however, significantly below unity, which indicates that the different scores measure different aspects of the test construct, using different measurement methods, and different sets of responses.

7.2.4 Differentiation among Known Populations

The next validity analysis examined whether or not VEPT Overall scores reflect expected differences between English language learners and native English speakers. Overall scores from 30 native speakers and 209 English language learners coming from a variety of first languages were compared. Figure 3 presents cumulative distributions of Overall scores for the learners and native English speakers.

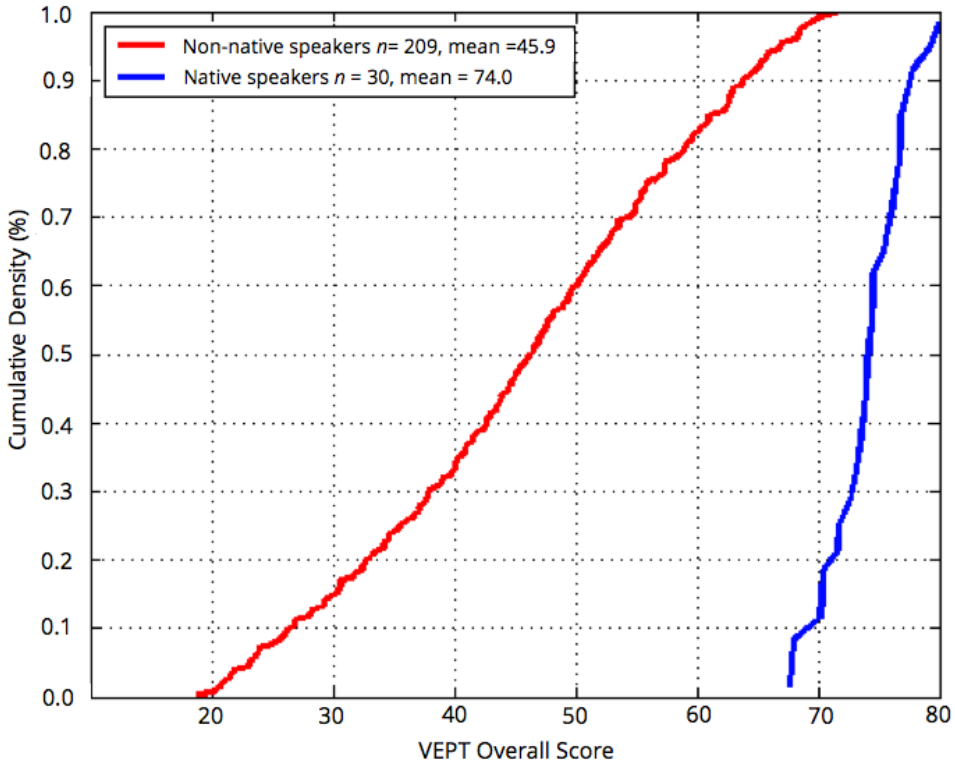


Figure 3. Cumulative density functions of VEPT Overall scores for native English speakers (n = 30) and English language learners (n = 209).

The results show that native speakers of English consistently obtain high scores on the VEPT. None of the native English speakers scored below 68. L2 speakers of English, on the other hand, are distributed over a wide range of scores. Note that only 3% of the English learners scored above 68. The Overall scores show effective separation between native English speakers and English learners.

7.3 Concurrent Validity

7.3.1 VEPT and IELTS

A study was conducted to understand the score relationship between the VEPT and the academic version of the International English Language Testing System (IELTS). IELTS is published and administered by

Cambridge ESOL. It is also designed to measure general English proficiency, in academic settings, on four language skills (Listening, Speaking, Reading, and Writing). IELTS takes approximately 2 hours 45 minutes to complete.

A group of 104 EFL learners at two academic institutions in Australia took both VEPT and IELTS within two months of each other. The learners were from 15 different countries, including China, Iran, Saudi Arabia, and Taiwan, representing a variety of mother tongues and ethnic backgrounds, with an equal balance of both genders and varying levels of English proficiency. All participants provided their official IELTS scores, the accuracy of which was verified by both academic institutions.

VEPT Overall scores and IELTS scores correlated at .87. Then, a simple regression was applied to the two sets of scores. Using the linear equation generated from the regression analysis, the VEPT Overall scores were translated into IELTS scores and a concordance table was created as shown in Table 10.

Table 10. Concordance table for VEPT Overall scores and IELTS Overall band scores

VEPT Overall Score	IELTS Overall Band
20-22	2.5
23-26	3.0
27-31	3.5
32-36	4.0
37-40	4.5
41-45	5.0
46-50	5.5
51-55	6.0
56-60	6.5
61-65	7.0
66-69	7.5
70-74	8.0
75-79	8.5
80	9.0

7.3.2 VEPT and TOEFL

Data were provided to Pearson for this study by a global education organization for 59 male and female EFL learners who had available scores on both the VEPT and Test of English as a Foreign Language (TOEFL) test. The participants represented a diverse sample of the population and had a variety of first langauges and ethnic backgrounds, as well as varying levels of English proficiency. About half (29) of the students' available TOEFL scores were from the TOEFL iBT (internet-based test), and the others (30) came from the TOEFL PBT (paper-based test). On its website ETS provides a score equivalence table for these two versions of TOEFL (http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf), which was used to translate scores from the PBT onto the iBT scale (0-120). The VEPT is reported on a 20-80 point scale.

TOEFL iBT takes approximately four hours and consists of integrated tasks that assess listening, speaking, reading, and writing. TOEFL PBT takes approximately two and a half hours, and it has three multiple choice sections (Listening Comprehension, Structure and Written Expression, and Reading Comprehension). VEPT takes approximately 50 minutes, assesses listening, speaking, reading, and writing, and is scored automatically by computer.

Some issues with the data provided to Pearson make this study problematic, however. As already mentioned, all students did not take the same two kinds of TOEFL test. Additionally, the TOEFL and VEPT test administrations did not coincide closely in time; in some cases there were many months between when students took the TOEFL and when they took VEPT. Finally, the range of scores suggested that the majority of data were from students of lower and middle proficiency levels, with only one TOEFL score at or above 100. With these caveats, a linear regression modeling technique was used to associate scores on these two tests using the available data. The observed correlation between VEPT and TOEFL scores was .85.

In light of the uneven distribution of data in the study, the relationship between the tests was smoothed to reflect the published cut scores on each test and the CEFR levels (Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015). Table 11 shows the adjusted correspondence between VEPT Overall Scores and TOEFL iBT Total Scores.

Table 11. Concordance table for VEPT Overall scores and TOEFL Overall band scores

VEPT Overall Score	TOEFL iBT Total Score
20	-
21	-
22	-
23	-
24	-
25	-
26	-
27	2
28	4
29	6
30	8
31	10
32	12
33	14
34	17
35	19
36	21
37	23
38	25
39	27

VEPT Overall Score	TOEFL iBT Total Score
51	57
52	60
53	63
54	65
55	68
56	70
57	73
58	75
59	77
60	80
61	82
62	84
63	86
64	88
65	90
66	92
67	94
68	96
69	98
70	99

40	29
41	31
42	33
43	35
44	38
45	40
46	43
47	46
48	49
49	52
50	55

71	101
72	103
73	104
74	106
75	107
76	109
77	110
78	112
79	113
80	114+

8. Linking to the Common European Framework of Reference for Languages

The Centre for Research in English Language Learning and Assessment at the University of Bedfordshire, in collaboration with Pearson, conducted a study to understand the relationship between the scores on the VEPT and the six levels of the Common European Framework of Reference for Languages (CEFR). The CEFR is published by the Council of Europe, and provides a common basis for describing language proficiency using a six-level scale: A1, A2, B1, B2, C1, and C2. This study included a standard setting procedure 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 potentially covered in the VEPT. This exercise was reviewed individually by six consultants at the University of Bedfordshire. After a review of the specification results, and receiving some CEFR familiarization training, a group of expert 14 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-centered 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 eight candidates' performances on the test as a whole.

Because 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. Many facet Rasch analysis (Linacre, 2015) 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). Because there was a sparse amount of data at the upper end the VEPT cut scores suggested by the standard setting procedure for these levels were not used. Instead, the Versant English Test and the Versant Writing Test cut scores were used. The results from previous studies mapping Versant English Test and Versant Writing Test scores onto the CEFR levels can be applied to VEPT because the tests share many tasks, the rating criteria are the same, the test items are linked through Item Response Theory psychometric modeling, and the underlying scoring technology is the same. Detailed reports on the two previous studies on Versant English Test and Versant Writing Test are available from Pearson on request. The results are summarized in Table 12.

Table 12. Mapping of CEFR Levels with VEPT scores

VEPT 20 – 80	CEFR <A1 – C2
20-23	<A1
24-33	A1
34-45	A2
46-56	B1
57-67	B2
68-78	C1
79-80	C2

9. Conclusions

This report has provided validity evidence in order to assist score users to make an informed interpretive judgment as to whether or not VEPT 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 VEPT scores are structurally reliable indications of candidate ability in spoken and written English and are suitable for decision-making.

10. 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. The VEPT 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).

11. References

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

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

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

Equal Employment Opportunity Commission. Uniform Guidelines for Employee Selection Procedures. Retrieved from http://www.uniformguidelines.com/uniformguidelines.html

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.

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*(2-3), 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.

Linacre, J. M. (2015) *Facets computer program for many-facet Rasch measurement*, version 3.71.4. Beaverton, OR: Winsteps.com

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

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*(8), 639-646.

Miller, G. A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2*(3), 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.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

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.

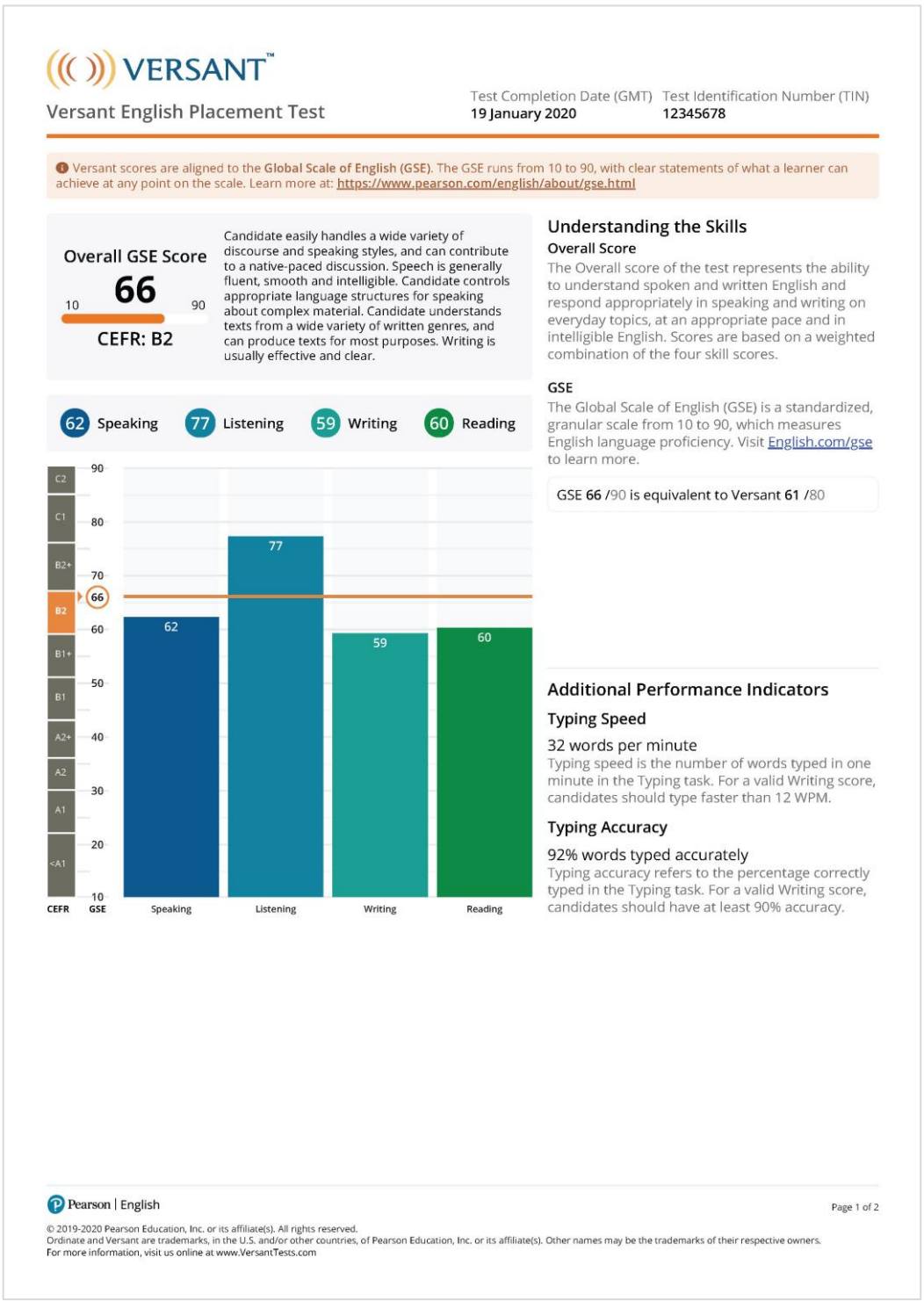
Sigott, G. (2004). *Towards identifying the C-test construct.* New York: Peter Lang.

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*(5363), 572-574.

12. Appendix

Side 1 of the Score Report: Summary of the candidate's Overall score and subscores.



VERSANT™

Versant English Placement Test

Test Completion Date (GMT) 19 January 2020

Test Identification Number (TIN) 12345678

Versant scores are aligned to the **Global Scale of English (GSE)**. The GSE runs from 10 to 90, with clear statements of what a learner can achieve at any point on the scale. Learn more at: https://www.pearson.com/english/about/gse.html

Overall GSE Score

66

10 — 90

CEFR: B2

Candidate easily handles a wide variety of discourse and speaking styles, and can contribute to a native-paced discussion. Speech is generally fluent, smooth and intelligible. Candidate controls appropriate language structures for speaking about complex material. Candidate understands texts from a wide variety of written genres, and can produce texts for most purposes. Writing is usually effective and clear.

62 Speaking | 77 Listening | 59 Writing | 60 Reading

Understanding the Skills

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 an appropriate pace and in intelligible English. Scores are based on a weighted combination of the four skill scores.

GSE

The Global Scale of English (GSE) is a standardized, granular scale from 10 to 90, which measures English language proficiency. Visit English.com/gse to learn more.

GSE **66** /90 is equivalent to Versant **61** /80

Additional Performance Indicators

Typing Speed

32 words per minute

Typing speed is the number of words typed in one minute in the Typing task. For a valid Writing score, candidates should type faster than 12 WPM.

Typing Accuracy

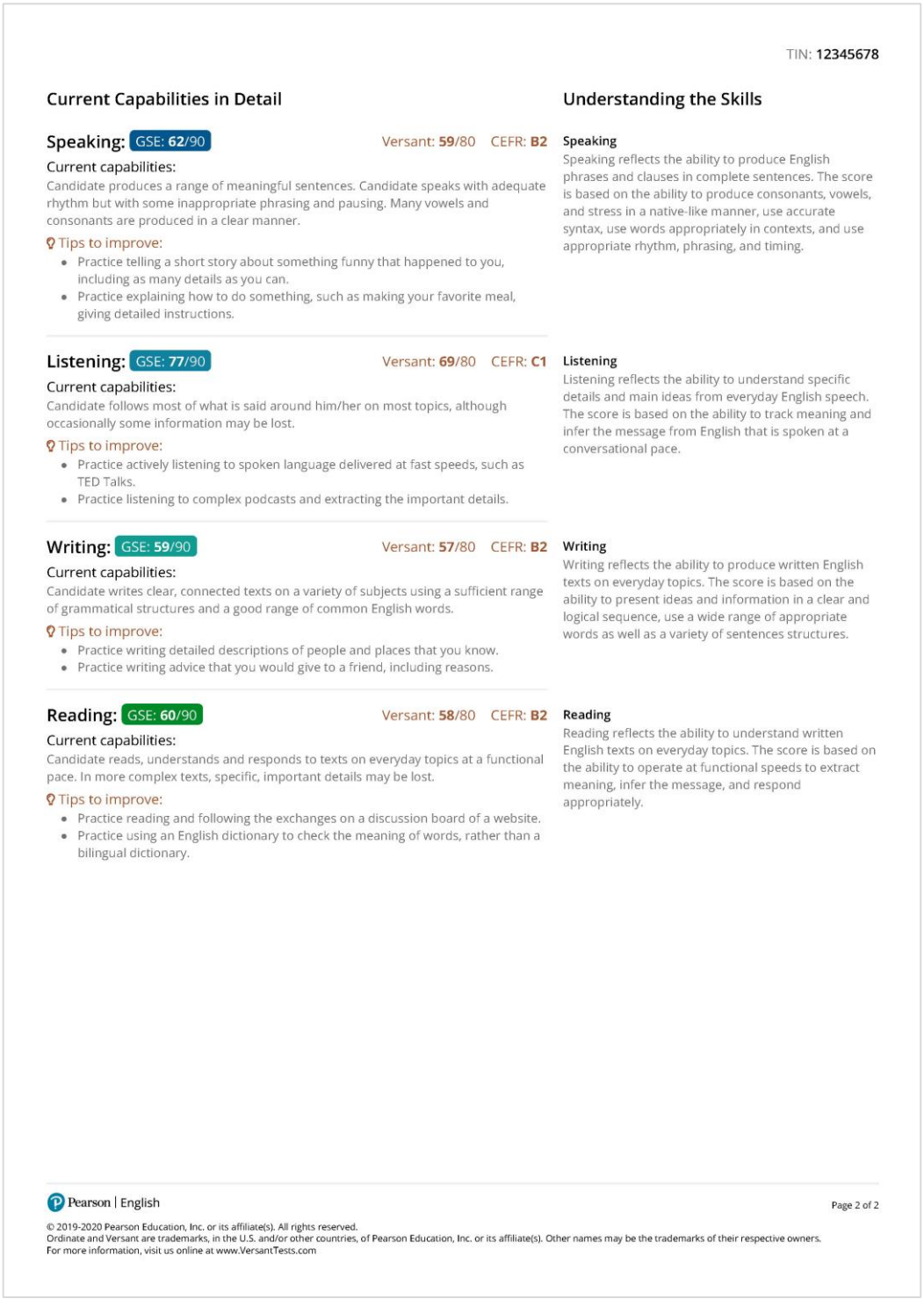
92% words typed accurately

Typing accuracy refers to the percentage correctly typed in the Typing task. For a valid Writing score, candidates should have at least 90% accuracy.

Pearson | English

Page 1 of 2

Side 2 of the Score Report: Detailed explanations of the candidate's language capabilities.



TIN: **12345678**

Current Capabilities in Detail

Speaking: GSE: **62**/90 Versant: **59**/80 CEFR: **B2**

Current capabilities:

Candidate produces a range of meaningful sentences. Candidate speaks with adequate rhythm but with some inappropriate phrasing and pausing. Many vowels and consonants are produced in a clear manner.

Tips to improve:

- Practice telling a short story about something funny that happened to you, including as many details as you can.
- Practice explaining how to do something, such as making your favorite meal, giving detailed instructions.

Listening: GSE: **77**/90 Versant: **69**/80 CEFR: **C1**

Current capabilities:

Candidate follows most of what is said around him/her on most topics, although occasionally some information may be lost.

Tips to improve:

- Practice actively listening to spoken language delivered at fast speeds, such as TED Talks.
- Practice listening to complex podcasts and extracting the important details.

Writing: GSE: **59**/90 Versant: **57**/80 CEFR: **B2**

Current capabilities:

Candidate writes clear, connected texts on a variety of subjects using a sufficient range of grammatical structures and a good range of common English words.

Tips to improve:

- Practice writing detailed descriptions of people and places that you know.
- Practice writing advice that you would give to a friend, including reasons.

Reading: GSE: **60**/90 Versant: **58**/80 CEFR: **B2**

Current capabilities:

Candidate reads, understands and responds to texts on everyday topics at a functional pace. In more complex texts, specific, important details may be lost.

Tips to improve:

- Practice reading and following the exchanges on a discussion board of a website.
- Practice using an English dictionary to check the meaning of words, rather than a bilingual dictionary.

Understanding the Skills

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 syntax, use words appropriately in contexts, and 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.

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.

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.

Pearson | English

Page 2 of 2

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