VERSANT™

Versant™ 4 Skills Essential Test

Test Description and Validation Summary

Table of Contents

1. Introduction 2

2. Test Description 2

2.1 Test Design 2

2.2 Test Administration 2

2.3 Test Format 3

2.4 Number of Items 7

2.5 Test Construct 7

2.5.1 Facility in Spoken and Written English 7

2.5.2 The Role of Memory 10

2.5.3 Context Independence 10

3. Content Design and Development 11

3.1 Vocabulary Selection 11

3.2 Item Development 11

3.3 Item Prompt Recording 12

3.3.1 Voice Distribution 12

3.3.2 Recording Review 12

4. Score Reporting 13

4.1 Scores and Weights 13

4.2 Score Use 15

5. Validation 15

5.1 Validity Study Design 16

5.2 Internal Validity 16

5.2.1 Descriptive Statistics 16

5.2.2 Standard Error of Measurement 16

5.2.3 Test Reliability 17

5.2.4 Dimensionality: Correlations Among Skill Scores 17

5.2.5 Machine Accuracy 18

5.3 Differentiation Among Known Populations 18

5.4 Linking to the Common European Framework of Reference for Languages 19

6. Conclusions 21

7. About the Company 21

8. References 22

1. Introduction

The Versant 4 Skills Essential Test, powered by Ordinate technology, is a web-delivered assessment instrument designed to measure how well a person can handle English on everyday and workplace topics. The Versant 4 Skills Essential Test is intended for adults over the age of 18 and takes approximately 30 minutes to complete. Because the Versant 4 Skills Essential Test 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 allows for immediate, objective, and reliable results that correspond well with traditional measures of spoken and written English performance.

The Versant 4 Skills Essential Test measures *facility* in spoken and written English. Facility in spoken and written English is how well a person can understand spoken and written English and respond appropriately in speaking and writing on everyday and business topics, at a native-like pace in intelligible English. Scores from the Versant 4 Skills Essential Test provide quick and reliable information that can be used for making decisions on recruitment in commercial and business organizations.

2. Test Description

2.1 Test Design

The Versant 4 Skills Essential Test has six tasks: Repeats, Sentence Builds, Conversations, Sentence Completion, Dictation, and Passage Reconstruction. These tasks provide multiple, fully independent measures that underlie facility in spoken and written English, including pronunciation, fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, 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 Versant 4 Skills Essential Test 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 facility in spoken and written English.

2.2 Test Administration

The Versant 4 Skills Essential Test is administered using Versant for Web (VfW), a browser-based test delivery application used on personal computers (see http://versantforweb.pearsontestservices.com for technical requirements). The Versant 4 Skills Essential Test can be taken at any time, from any location. Due to the automated administration, a human examiner is not required.

Administration of the Versant 4 Skills Essential Test generally takes about 30 minutes. During test administration, an examiner's voice guides the candidate through the test, explains the tasks, and gives examples and instructions. The candidate also listens through a headset and sees the instructions and examples on the computer screen. Candidates respond to test questions by speaking into the microphone or by typing a response.

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 items, the candidate has a set amount of time to respond. 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.

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 and posts scores to a secure website, usually within minutes of completing the test. Test administrators and score users can view and print out test results from ScoreKeeper, a password-protected section of Pearson's website (www.pearson.com/versant).

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 six parts of the Versant 4 Skills Essential Test.

Part A: Repeats

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 3 to 15 words. The audio item prompts are spoken in a conversational manner.

Examples:

1. He's a great manager.
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 the words as spoken in a continuous stream of speech (Miller & Isard, 1963). Highly proficient speakers of English can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with English words, phrase structures, and other common syntactic forms. If a person habitually processes five-word phrases as a chunk (e.g. "the really big apple tree"), then that person is capable of repeating 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, Repeat 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 B: 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 random order (excluding the original word order), and the candidate responds by saying a grammatical sentence made up of a re-arrangement of the three given phrases.

Examples:

1. my boss / to London / moved
2. the prices range / to thirty dollars / from fifteen
3. to their leader / listened carefully / the young men

To correctly complete this task, a candidate must understand the possible meanings of the phrases and know how they might combine with other phrasal material, both with regard to syntax and pragmatics. The length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one-word versus a three-word phrase) that a person can hold in verbal working memory. This is important to measure because it reflects the candidate's ability to access and retrieve lexical items and to build phrases and clause structures automatically. The more automatic these processes are, the more the 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 entire sentences. As such, it is a measure of candidates' mastery of sentences in addition to their pronunciation and fluency.

Part C: Conversations

In the Conversations task, candidates listen to a conversation between two speakers, which typically consists of three short sentences. Immediately after the conversation, an examiner voice asks a comprehension question and candidates answer the question with a word or short phrase.

Example:

Speaker 1: How was your business trip?
Speaker 2: There were thunderstorms the whole time.
Speaker 1: That sounds terrible.

Question: What happened during the business trip?

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 question.

Part D: Sentence Completion

In this task, candidates read a sentence that has a word missing, and 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, retrieve a lexical item to complete the sentence, and type the word in the text box provided. Sentences range in length from 5 to 25 words. Across all items in this task, candidates are exposed to sentences with words missing from various parts of speech (e.g., noun, verb, adjective, adverb) and from different positions in sentences: sentence-initial, sentence-medial, and 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 (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."

In many Sentence Completion items there is more than one possible correct answer choice. Based on responses from field testing, all reasonable correct answer choices are accepted. 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 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 E: Dictation

In the Dictation task, candidates hear a sentence and they must type the sentence 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 words to 14 words. The items present a range of grammatical and syntactic structures, including imperatives, *wh*-questions, contractions, plurals, possessives, various tenses, and particles. The audio item prompts are spoken with a natural pace and rhythm by various native and non-native speaker voices that are distinct from the examiner voice.

Examples:

1. There's hardly any paper left.
2. Success is impossible without teamwork.
3. 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 memory or to reconstruct the sentence from the memory traces that are available. 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 is a good test of comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for 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: Passage Reconstruction

Passage Reconstruction is similar to a task known as free recall. Candidates are asked to read a text, put it aside, and then write what they can remember from the text in a limited amount of time. 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, or results.

In order to accurately reconstruct a passage, 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 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, then the main ability assessed is reading comprehension, but when the passage is reconstructed in the target language (in this case, English), then it is more an integrated test of both reading and writing (Alderson, 2000, p. 230).

2.4 Number of Items

In the administration of the Versant 4 Skills Essential Test, the testing system serially presents a total of 70 items in 6 separate tasks to each candidate. Table 1 shows the number of items presented in each task.

Table 1. Number of items presented per task

Task	Number of Items
A. Repeats	16
B. Sentence Builds	8
C. Conversations	12
D. Sentence Completion	18
E. Dictation	14
F. Passage Reconstruction	2
Total	**70**

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 Versant 4 Skills Essential Test is designed to measure a candidate's facility in spoken and written English—that is, how well a person can understand spoken and written English and respond appropriately in speaking and writing on everyday and workplace topics, at a native-like pace and in intelligible English.

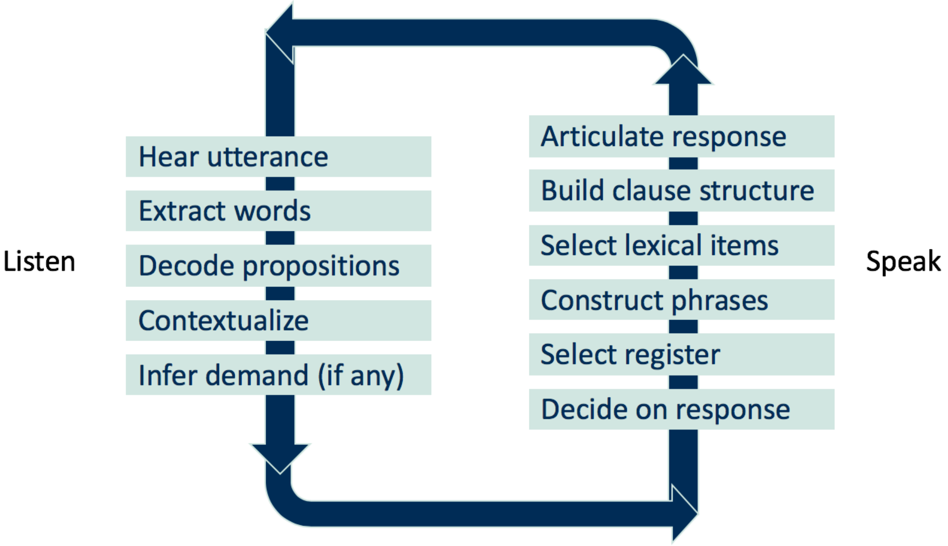
The first concept embodied in the definition of facility is how well a candidate *understands* spoken and written English. Both receptive modalities (listening and reading) are used in the test. Repeats, Sentence Builds, Conversations, and Dictation expose candidates to spoken English, and Sentence Completion and Passage Reconstruction 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 and Passage Reconstruction require fluent word recognition and problem-solving comprehension abilities (Carver, 1991). Interestingly, the initial and simplest step in the reading process—word recognition— is what 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 Versant 4 Skills Essential Test 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).



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 task, Passage Reconstruction, is designed to assess not only proficiency in the core linguistic skills of grammatical and lexical range and accuracy, but also the other essential elements of good writing such as coherence and cohesion. 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 and highlighting salient points with discourse markers.

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 Versant 4 Skills Essential Test probes the degree of automaticity in language performance.

By utilizing integrated tasks, the Versant 4 Skills Essential Test 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 Versant 4 Skills Essential Test 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 primarily memory tests. Because some Versant 4 Skills Essential Test 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. Memory is the primary component of intelligence and linguistic abilities; it is important to be sure that items do not tap out memory capacity, but just use what is considered amounts of it within all people's normal range.

2.5.3 Context Independence

The Versant 4 Skills Essential Test probes the psycholinguistic elements of spoken and written language performance rather than the social, rhetorical, and cognitive elements of communication. All items 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 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 Versant 4 Skills Essential Test are knowledge of the language, such as grammar and vocabulary, comprehension of the information conveyed through the language, knowledge of 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 that are common in everyday English. This includes extensive use of pronominal expressions such as "she" or "their friend" and contracted forms such as "won't" and "I'm." 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

Versant 4 Skills Essential Test items were drafted by trained 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 and workplace settings. 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
- Fax/Telephone/E-Mail
- Inventory
- Scheduling
- Marketing/Sales

Item writers were asked to write items that are independent of social nuance and complex cognitive functions, and that 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 reviewers 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, including anticipated responses for Conversation and 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 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

Twenty-nine native speakers (14 women and 15 men) 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 whose country of origin included India, Korea, and Costa Rica. 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.

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 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 insertion in the operational test.

4. Score Reporting

4.1 Scores and Weights

The Versant 4 Skills Essential Test 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 and workplace 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 20 to 80.

Speaking: Speaking reflects the ability to produce intelligible communication in spoken English in everyday and workplace situations. 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 and workplace 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 and workplace topics. The score is based on the ability to operate at functional speeds to extract details and main ideas, infer the message, and construct meaning.

Writing: Writing reflects the ability to produce accurate, appropriate written responses at a functional pace on everyday and workplace 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, use appropriate register, as well as a variety of sentence structures.

From the 70 items on a Versant 4 Skills Essential Test form, 66 responses are used in the automatic scoring. The first item in Parts A, B, C, and E is considered a practice item and is not incorporated into the final score. Figure 2 illustrates which tasks of the test contribute to each of the four skill scores. Each vertical rectangle represents a response from a candidate. The items that are not included in the automatic scoring are shown in blue. Figure 2 illustrates which tasks of the test contribute to each of the four skill scores.

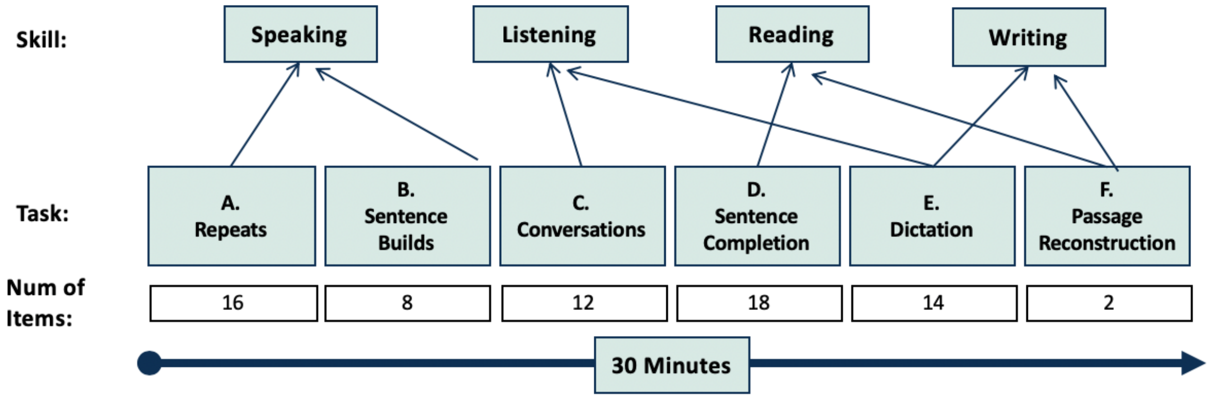


Figure 2. Relation of skill scores to tasks

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

Table 2. Skill score weighting in relation to Versant 4 Skills Essential Test Overall score

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

In the Versant 4 Skills Essential Test 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 Versant 4 Skills Essential Test 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 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 Versant 4 Skills Essential Test 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 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 are typically business organizations and academic programs. 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 Versant 4 Skills Essential Test themselves under secure conditions, or by having trusted third parties administer the test. In several countries, education and commercial institutions provide such services.

Versant 4 Skills Essential Test scores can be used to assess how well and efficiently a candidate can process and produce spoken and written English on everyday and workplace topics. The Versant 4 Skills Essential Test score scale covers a wide range of abilities in spoken and written English communication; therefore, it is effective for recruiting environments that need quick and reliable results.

It is up to score users to decide what Versant 4 Skills Essential Test 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. Validation

Automated scoring methods are used for both the spoken and written responses in the Versant 4 Skills Essential Test. The scoring models used in the Versant 4 Skills Essential Test were trained on a norming data set comprised of a large number of native and non-native English-speaking test-takers. Test-takers came from a wide variety of different countries, including Indonesia, Singapore, China, Japan, Korea, and India. 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, crucial to validate the machine-generated scores on a new set of data.

5.1 Validity Study Design

Validity analyses demonstrate three aspects of the Versant 4 Skills Essential Test scores:

- Internal quality: whether or not the Versant 4 Skills Essential Test a) provides scores that are reliable and internally consistent, b) provides distinct subscores that measure different aspects of the test construct, and c) provides scores that are comparable to the scores that human listeners and raters assign
- Relation to known populations: whether or not the Versant 4 Skills Essential Test scores reflect expected differences and similarities among known populations (e.g., natives vs. English learners)
- Relation to scores of tests with related constructs: how do Versant 4 Skills Essential Test scores correspond to the six levels of the Common European Framework of Reference (CEFR).

For the reliability and validation analyses, a total of 2851 test takers who took the Versant 4 Skills Essential Test between January 2019 and June 2019 were selected. Over 20 different first language backgrounds were represented in the validation sample. Ages ranged from 18 to 65 (average of 29) and the male:female ratio was 57:43. Care was taken to ensure that the training dataset and validation dataset did not overlap for independent validation 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.

5.2 Internal Validity

5.2.1 Descriptive Statistics

The mean Overall score of the validation sample was 56.15 with a standard deviation of 14.16 (on a scale of 20-80). Table 3 summarizes some descriptive statistics for the validation sample.

Table 3. Descriptive Statistics for the Validation Dataset (*N*=2851)

Measure	Statistic
Mean	56.15
Standard Error	0.27
Median	55.00
Standard Deviation	14.16

5.2.2 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). 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 Versant 4 Skills Essential Overall score is 2.11. In other words, if a candidate received an Overall score of 50 on Versant 4 Skills Essential Test, we are 96% confident that this person's "true" overall score falls between 45.8 and 54.2.

5.2.3 Test Reliability

A separate analysis was not conducted for the Versant 4 Skills Essential Test with regard to test reliability. Thus, split-half reliability calculated for the Versant English Placement Test (VEPT) (Pearson, 2019) is referred to here instead given that (a) VEPT is another similar Versant test assessing 4 different skills (speaking, listening, reading, and writing) (b) the two tests share many item types in common (all the 6 item types in the Versant 4 Skills Essential Test are present in VEPT), and (c) the two tests use the same automated scoring models.

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 on VEPT. Table 4 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 4. 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 4 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. The same arguments can be made for the Versant 4 Skills Essential Test based on the similarities between VEPT and Versant 4 Skills Essential Test.

5.2.4 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 2581 Versant 4 Skills Essential Tests delivered over a six-month period was used for the analysis. Table 5 presents the correlations among the Versant 4 Skills Essential Test scores for this sample.

Table 5. Inter-correlation between skill scores on the Versant 4 Skills Essential Test (*N*=2581)

	Speaking	Listening	Reading	Writing	Overall
Speaking	-	.81	.66	.74	**.90**
Listening		-	.71	.76	**.91**
Reading			-	.87	**.88**
Writing				-	**.92**

As expected, skill scores correlate with each other to a moderate extent 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, which indicates that the different scores measure different aspects of the test construct, using different measurement methods, and different sets of responses.

5.2.5 Machine Accuracy

Another analysis for internal quality would involve comparing scores from the Versant 4 Skills Essential Test, which uses automated language processing technologies, with scores derived from human transcriptions and human judgments by expert raters. Since a separate analysis was not conducted in this regard for the Versant 4 Skills Essential Test, an analysis of VEPT (Pearson, 2019) is presented here given the similarities between the two tests mentioned earlier.

Table 6 presents correlations between machine scores and human scores. Correlations presented in Table 6 suggest that scoring a VEPT by machine yields scores that closely correspond with human ratings.

Table 6. Correlations between human and machine scoring of VEPT responses (*N*=214)

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

The values in Table 10 suggest that the human-machine relation is close for all four skill scores, and 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. Considering the close resemblance between VEPT and Versant 4 Skills Essential Test in terms of item types and machine scoring models, the human-machine correlation for Versant 4 Skills Essential Test is expected to be of similar magnitude.

5.3 Differentiation Among Known Populations

Another piece of validity evidence from VEPT, which is closely comparable with Versant 4 Skills Essential Test, shows that Overall scores reflect expected differences between English language learners and native English speakers (Pearson, 2019). 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.

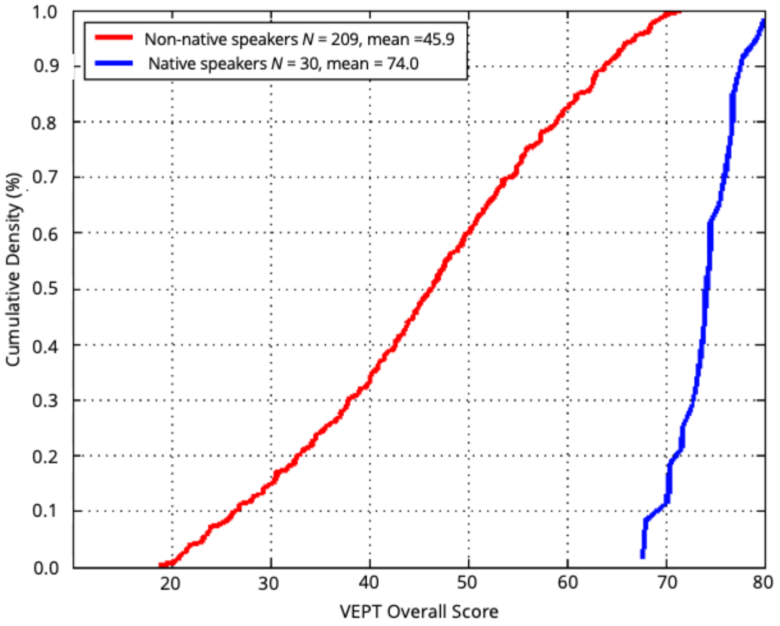


Figure 3. Cumulative density functions of VEPT Overall scores for native English speakers (*N*=30) and English language learners (*N*=209).

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 demonstrate effective separation between native English speakers and English learners. Assuming the close resemblance between VEPT and Versant 4 Skills Essential Test in terms of item types and scoring system, the scores on Versant 4 Skills Essential Test are also expected to reliably distinguish native English speakers and English learners.

5.4 Linking to the Common European Framework of Reference for Languages

The score report of Versant 4 Skills Essential Test provides an estimated CEFR level corresponding to the Versant Overall score that a candidate has received, and the linking is based on the previous study mapping VEPT scores onto the CEFR levels (Green, 2013). The result from the previous study can be applied to the Versant 4 Skills Essential Test because the tests share many tasks, the rating criteria are the same, the test items are linked through Item Response Theory statistical modeling, and the underlying scoring technology is the same.

According to Green (2013), 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 (Council of Europe, 2001). 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. Multi-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). The results are summarized in Table 7.

Table 7. 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

6. Conclusions

Data from the validation studies provide evidence in support of the following conclusions:

- The Versant 4 Skills Essential Test produces precise and reliable skill estimates.
- Overall scores show effective separation between native and non-native candidates.
- Subscores of the Versant 4 Skills Essential Test are reasonably distinct and therefore offer useful diagnostics.
- Versant 4 Skills Essential Test scores show a high correlation with human-produced ratings.
- Versant 4 Skills Essential Test Overall scores correspond to the CEFR levels.

With valid, reliable automatic scoring that is virtually indistinguishable from human scorers, the 30-minute web-delivered Versant 4 Skills Essential Test provides an ideal solution for recruiting environments that need quick and trustworthy results.

7. About the Company

Pearson: Pearson's Knowledge Technologies group and Ordinate Corporation, the creator of the Versant tests, were combined in January 2008. The group is currently a part of the Assessment Technology Engineering group within Pearson. The Versant tests were the first to leverage a completely automated method for assessing spoken and written language.

Ordinate 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 telephone 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 Ordinate technology. Research results are published in international journals and made available through the Versant website (www.VersantTests.com).

8. 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.

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.

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.

Green, A. (2013). Relating the Versant English Placement Test to the Common European Framework of Reference. Centre for Research in English Language Learning and Assessment. University of Bedfordshire.

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.

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: 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.), *Hoi̇̈rverstäi̇ndnis 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. (2019). Versant English Placement Test: Test description and validation summary. Unpublished manuscript.

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

Pearson creates unique technology for automated assessment of speech and text used in a variety of industry leading products and services. These include the Versant line of automated spoken language tests built on Ordinate technology, and WriteToLearn™ automated written summary and essay evaluations using the Knowledge Analysis Technologies™ (KAT) engine.

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

www.VersantTests.com

Version 0719A