



# **Versant™ Pro - Speaking Test**

**Test Description and Validation Summary**

# Table of Contents

<b>1. Introduction</b>	<b>4</b>
<b>2. Test Description</b>	<b>4</b>
2.1 Workplace Emphasis	4
2.2 Test Design	4
2.3 Test Administration	5
2.3.1 Telephone Administration	5
2.3.2 Computer Administration	5
2.4 Test Format	6
Part A: Read Aloud	6
Part B: Repeats	7
Part C: Short Answer Questions	7
Part D: Sentence Builds	8
Part E: Story Retelling	8
Part F: Response Selection	9
Part G: Conversations	9
Part H: Passage Comprehension	9
2.5 Number of Items	10
2.6 Test Construct	11
<b>3. Content Design and Development</b>	<b>12</b>
3.1 Rationale	12
3.2 Vocabulary Selection	13
3.3 Item Development	13
3.4 Item Prompt Recordings	14
3.4.1 Item Recording	14
3.4.2 Recording Review	14
<b>4. Score Reporting</b>	<b>14</b>
4.1 Scoring and Weighting	14
4.2 Score Use	16
<b>5. Field Test</b>	<b>17</b>
5.1 Data Collection	17
5.1.1 Native Speakers	17
5.1.2 Non-Native Speakers	17
<b>6. Data Resources for Scoring Development</b>	<b>18</b>
6.1 Data Preparation	18
6.1.1 Transcription	18
6.1.2 Expert Human Rating	18
<b>7. Validation</b>	<b>19</b>
7.1 Validity Study Design	19
7.1.1 Validation Sample	19
7.2 Internal Validity	20
7.2.1 Descriptive Statistics	20

7.2.2 Test Reliability .....	20
7.2.3 Dimensionality: Correlations Among Subscores.....	21
7.2.4 Machine Accuracy.....	23
7.2.5 Differentiation among Known Populations .....	23
7.3 Concurrent Validity .....	24
7.3.1 Versant Pro – Speaking and TOEIC.....	24
7.3.2 Versant Pro – Speaking and Versant English Test.....	26
7.3.3 Versant Pro – Speaking and CEFR Level Estimates .....	28
<b>8. Conclusion .....</b>	<b>28</b>
<b>9. About the Company .....</b>	<b>29</b>
<b>10. References.....</b>	<b>29</b>
<b>11. Appendix A: Test Paper .....</b>	<b>31</b>

## 1. Introduction

Pearson's Versant™ Pro – Speaking Test, powered by Ordinate technology, is a telephone-based or computer-based assessment instrument, which is designed to evaluate how well a person can understand and speak English at a conversational pace on everyday and workplace topics. Versant Pro – Speaking is intended for adults 18 years of age and older and takes about 25 minutes to complete. Because the Versant Pro – Speaking test is delivered automatically by the Versant testing system, the test can be taken at any time, from any location. A human examiner is not required. The computerized scoring allows for immediate and objective results that are reliable and correspond well with traditional measures of English language proficiency.

Versant Pro – Speaking measures *facility* with spoken English in the workplace context, which is *how well a person can understand spoken English and respond appropriately on everyday and workplace topics at a native-like conversational pace*. Versant Pro - Speaking scores provide reliable information that can be applied to placement, qualification and certification decisions by academic institutions, businesses and government agencies. The test is also appropriate for monitoring progress as well as measuring instructional outcomes. (Versant Pro – Writing<sup>1</sup> is also available if written English is important in decision-making.)

## 2. Test Description

### 2.1 Workplace Emphasis

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

### 2.2 Test Design

Versant Pro – Speaking has eight item types or sections: Reading, Repeat, Short Answer Questions, Sentence Builds, Story Retelling, Response Selection, Conversations, and Passage Comprehension. All items in Versant Pro – Speaking elicit responses from the candidate that are analyzed automatically. These item types provide multiple, fully independent measures that underlie facility in spoken English, including oral reading fluency, phonological fluency, pronunciation of rhythmic and segmental units, sentence comprehension and construction, and passive and active vocabulary use. Because more than one item type contributes to each subscore, the use of multiple item types strengthens score reliability.

The Versant Pro – Speaking score report is comprised of an Overall score and five diagnostic subscores:

- Sentence Mastery
- Vocabulary
- Fluency
- Pronunciation
- Listening Comprehension

---

<sup>1</sup> For more information about Versant Pro – Writing, please refer to *Versant Pro – Writing: Test Description and Validation Summary*.

The Overall score is a weighted average of the five subscores. Together, these scores describe the candidate's facility in spoken English in everyday and workplace contexts.

The Versant testing system automatically analyzes the candidate's responses 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.VersantTest.com](http://www.VersantTest.com)).

## 2.3 Test Administration

Administration of a Versant Pro – Speaking test generally takes about 30 minutes over the phone or via a computer. It is best practice (even for computer delivered tests) for the administrator to give a test paper to the candidate at least five minutes before starting the test (see Appendix). The candidate then has the opportunity to read both the test instructions and the test paper and ask questions before the test begins. The administrator should answer any procedural or content questions that the candidate may have.

The delivery of the recorded item prompts is interactive – the system detects when the candidate has finished responding to one item and then presents the next item.

### 2.3.1 Telephone Administration

Telephone administration is supported by test instructions and a test paper. The test instructions contain general instructions and an explanation of the test procedures. These instructions are the same for all candidates. The test paper is a single sheet of paper with material printed on both sides. It is unique for each candidate. It contains the phone number to call, the Test Identification Number (TIN), the spoken instructions written out verbatim, item examples, and the printed sentences for Part A: Read Aloud.

When the candidate calls the Versant testing system, the system will ask the candidate to use the telephone keypad to enter the Test Identification Number that is printed on the test paper. This identification number is unique for each candidate and keeps the candidate's information secure.

A single examiner voice presents all the spoken instructions for the test. The spoken instructions for each section are also printed verbatim on the test paper to help ensure that candidates understand the directions. Candidates interact with the test system in English, going through all eight parts of the test until they complete the test and hang up the telephone.

### 2.3.2 Computer Administration

For computer administration, the computer must have an Internet connection and Pearson's Computer Delivered Test (CDT) software (available at <http://www.versanttest.com/technology/platforms/cdt/index.jsp>). The candidate is fitted with a microphone headset. The CDT software prompts the candidate to adjust the volume and calibrate the microphone before the test begins.

The instructions for each section are spoken by an examiner voice and are also displayed on the computer screen. Candidates interact with the test system in English, speaking their responses into the microphone. When a test is finished, the candidate clicks a button labeled, "End Test".

## 2.4 Test Format

Test items are spoken by various native English speakers including American, British and Australian as well as highly proficient non-native speakers. Voices of these test items are distinct from the examiner voice.

The following subsections provide brief descriptions of the item types and the abilities required to respond to the items in each of the eight parts of the Versant Pro – Speaking Test.

### 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 reading texts are printed on the test paper or displayed on the computer screen.

The passages take the form of either an expository text or an email message and deal with typical business topics or activities. All passages are relatively simple in structure and vocabulary and range in length from 40 to 55 words. The SMOG Readability Index (<http://www.harrymclaughlin.com/SMOG.htm>) was used to identify and refine the readability score for each passage. SMOG estimates the number of years of education needed to comprehend a passage. The algorithm factors in the number of polysyllabic words across sentence samples (McLaughlin, 1969). All passages have a readability score between 9 and 12, which is at a high school level. 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 page or 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.

Examples:

1. Many companies are becoming more and more diverse in the current global market. Some companies encourage diversity in their workplace. The key to a successful work environment is to appreciate each other's background. The goal is to embrace diversity rather than deny differences between people.
2. We have several offices for rent in a large office building. The building is surrounded by trees. All offices have private balconies and hardwood floors. There are many features including an outdoor eating area and a shower. The location is within a few steps of many shops and cafes.

## Part B: Repeats

In this task, candidates are asked to repeat sentences verbatim. The administration is interactive. The system plays a sentence spoken by a native speaker and the candidate attempts repeating it; then the system plays another sentence and the candidate repeats it. The interaction continues in this way until the candidate completes the section. The sentences are presented to the candidate in approximate order of increasing difficulty. Sentences range in length from 3 words to 15 words. The audio item prompts are spoken in a conversational manner.

Examples:

1. It took a lot longer than expected.
2. Come to my office after class if you need help.
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 unit (e.g. “the really big apple tree”), then that person can usually repeat utterances of 15 or 20 words in length. Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not familiar with English sentence structure.

Because the Repeat items require candidates to organize speech into linguistic units, 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 C: Short Answer Questions

In this task, candidates listen to spoken questions and answer each question with a single word or short phrase. The questions generally present at least three or four lexical items spoken in a continuous speech and framed in English sentence structure. Each question asks for basic information or requires simple inferences based on time, sequence, number, lexical content, or logic. The questions do not presume any knowledge of specific facts of culture, geography, history, or other subject matter; they are intended to be within the realm of familiarity of both a typical 12-year-old native speaker of English and an adult who has never lived in an English-speaking country.

Examples:

1. How many sides does a triangle have?
2. Would you be more likely to find a handle on a suitcase or a pen?
3. If Mark is having a breakfast meeting, is the meeting in the morning or in the afternoon?

To correctly respond to the questions, a candidate needs to identify the words in phonological and syntactic context, and then infer the demand proposition. Short Answer Questions measure receptive and productive vocabulary within the context of spoken questions presented in a conversational style.

## Part D: Sentence Builds

In 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 is expected to say a reasonable and grammatical sentence that comprises exactly the three given phrases.

Examples:

1. my boss / to California / 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 (see Section 2.6, Test Construct).

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 E: Story Retelling

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

Example:

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

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

## Part F: Response Selection

In the Response Selection task, candidates listen to a sentence, which is immediately followed by three possible responses. From among the three possible responses, candidates choose the one that is the most appropriate response to the sentence. Candidates answer each question either by clicking 'A', 'B', or 'C' on the computer in the case of CDT delivery or by saying 'A', 'B', or 'C' into the phone in the case of the telephone delivery.

Example:

Our profit last year was higher than expected.

A: Great, let's celebrate.

B: That's too bad.

C: We lost a lot last year.

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

## Part G: Conversations

In the Conversation 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 the business trip?

Speaker 2: There was a storm 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 H: Passage Comprehension

In the Passage Comprehension task, candidates listen to a spoken passage (usually a story) and then are presented with three comprehension questions about the passage. The passages range from 40 to 70 words in length. Most passages are simple stories with a situation involving a character (or characters), a setting, and an ending. The body of the story typically describes an action performed by the agent of the story followed by a possible reaction or implicit sequence of events. The ending typically introduces a result, new situation, actor, patient, thought, or emotion.

Example:

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

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

Question 1: What problem did Jason have when he woke up?

Question 2: What did he do right after calling his boss?

Question 3: What did Jason do that afternoon?

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

## 2.5 Number of Items

In total, 88 items are presented to each candidate in the eight separate sections, Parts A through H. In each task section, the items are drawn from a much larger item pool. For example, each candidate is presented with 16 Repeat items selected quasi-randomly from the pool. Most or all items will be different from one test administration to the next. The Versant testing system selects items from the item pool taking into consideration, among other things, the item's level of difficulty and its form and content in relation to other selected items. Table I shows the number of items presented in each section.

Table I. Number of items presented per section.

Task	Presented
A. Read Aloud	2
B. Repeats	16
C. Short Answer Questions	20
D. Sentence Builds	10
E. Story Retelling	3
F. Response Selection	16
G. Conversations	12
F. Passage Comprehension	9 (3 passages)
Total	88

## 2.6 Test Construct

For any language test, it is essential to define the test construct as explicitly as possible (Bachman, 1990; Bachman & Palmer, 1996). The Versant Pro – Speaking test is designed to measure a candidate's *facility in spoken English* – that is the ability to understand spoken English on everyday and workplace topics and to respond appropriately at a native-like conversational pace in intelligible English. Another way to express the construct, *facility in spoken English*, is “ease and immediacy in understanding and producing appropriate conversational English” (Levelt, 1989). There are many processing elements 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).

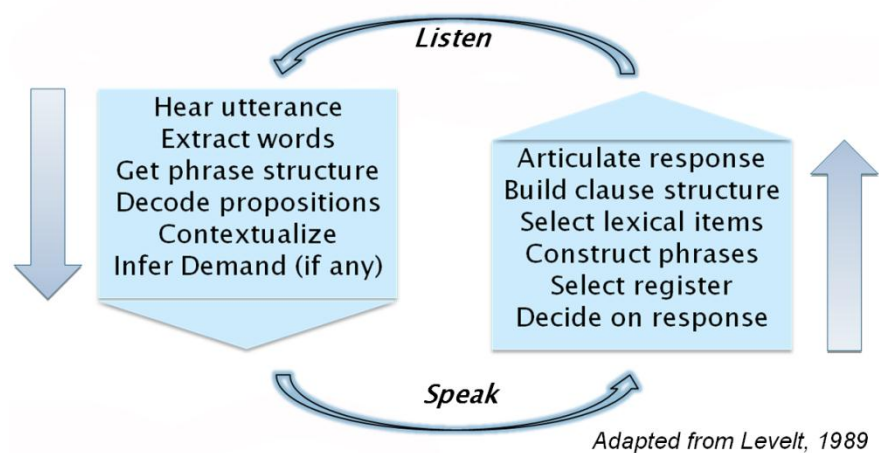


Figure 1. Conversational processing components in listening and speaking.

Core language component processes, such as lexical access and syntactic encoding, typically take place at a very rapid pace. During spoken conversation, Van Turenhout, Hagoort, and Brown (1998) found that 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 500-1000 milliseconds (Bull and 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.

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 speaker/listener 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, Versant Pro - Speaking probes the degree of automaticity in language performance.

Some measures of automaticity can be misconstrued as memory tests. Since some Versant Pro - Speaking tasks involve repeating long sentences or holding phrases in memory in order to assemble them into reasonable sentences, 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. Note that every Repeat and every Sentence Build item on the test was presented to a

sample of educated native speakers of English and at least 85% of the speakers in that educated native speaker 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. Also, if memory capacity (rather than language ability) were a principal component of the variation among people performing these tasks, the test would not correlate so closely with other accepted measures of oral proficiency (see Section 7, Validation).

Three basic types of scores are produced from the test: scores relating to the content of what a candidate says, scores relating to the manner of the candidate's speaking, and scores relating to the candidate's listening proficiency. For the speaking part of the scores (i.e., content and manner), this distinction corresponds roughly to Carroll's (1961) description of a knowledge aspect and a control aspect of language performance. In later publications, Carroll (1986) identified the control aspect as automatization, which occurs when speakers can talk fluently without realizing they are using their knowledge about a language.

The Versant Pro – Speaking test probes the psycholinguistic elements of spoken language performance rather than the social, rhetorical, and cognitive elements of communication. The reason for this focus is to ensure that test performance relates most closely to the candidate's facility with the language itself and is not confounded with other factors. The goal is to separate familiarity with spoken language from other types of knowledge including cultural familiarity, understanding of social relations and behavior, and the candidate's own cognitive style. Also, by focusing on context-independent material, less time is spent developing a background cognitive schema for the tasks, and more time is spent collecting data for language assessment.

The Versant Pro – Speaking test provides a measurement of the real-time decoding and encoding of spoken English. Performance on Versant Pro - Speaking items predicts a more general spoken English facility, which is essential for successful oral communication in English. The same facility in spoken English that enables a person to satisfactorily understand and respond to the listening/speaking tasks in Versant Pro - Speaking also enables that person to participate in native-paced conversation.

## 3. Content Design and Development

### 3.1 Rationale

All Versant Pro – Speaking item content is designed to be region-neutral. The content specification also requires that both native speakers and proficient non-native speakers find the items easy to understand and to respond to appropriately. For English learners, the items probe a broad range of skill levels and skill profiles.

Except for some of the listening items (i.e., Response Selection, Conversations, Passage Comprehension), all items are context independent, spoken material in English. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based (Perry, 2001). Second, when language usage is relatively context-independent, task performance depends *less* on factors such as world knowledge and cognitive style and *more* on the candidate's facility with the language itself. Thus, the test performance relates most closely to language abilities and is not confounded with other candidate characteristics. Third, context-independent tasks maximize response density; that is, within the time allotted for the test, the candidate has more time to

demonstrate performance in speaking the language because less time is spent presenting contexts that situate a language sample or set up a task demand.

### 3.2 Vocabulary Selection

The vocabulary used in the test items was taken from a general English corpus and a business English word list. The general English corpus was restricted to forms of the 8,000 most frequent words found in the Switchboard Corpus (Godfrey and Holliman, 1997), a corpus of three million words taken from spontaneous telephone conversations. The business English word list was restricted to forms of the 3,500 most frequent words found in the *University of Cambridge Business English Certificate Preliminary Wordlist*, *Barron's 600 Essential Words for the TOEIC*, and *Oxford Business and Finance words*.

### 3.3 Item Development

The Versant Pro – Speaking items were drafted by trained item writers. All the items writers have advanced degrees or training in applied linguistics, TESOL, or language testing. In general, structures used in the test reflect those that are used in common everyday or 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/Email
- Inventory
- Scheduling
- Marketing/Sales

Item writers were specifically requested to write items that would not favor candidates with work experience or require any work experience to answer correctly. The items were designed to be independent of social and cultural nuances, and high-cognitive functions. The items are intended to be within the realm of familiarity of both a typical 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 linguists for expert review to ensure 1) compliance with the vocabulary specification, and 2) conformity with current colloquial English usage in different countries. Reviewers checked that items would also be appropriate for candidates trained to standards other than American English.

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

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

## 3.4 Item Prompt Recordings

### 3.4.1 Item Recording

Thirty native speakers (14 women and 16 men) representing various speaking styles and regions, such as the U.S. U.K. and Australia, were selected for recording the spoken prompt materials.

Several non-native speakers also recorded some items. 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. The speakers' country of origin included India, Hong Kong, Taiwan, Korea, and the Netherlands.

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.

### 3.4.2 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 installation in the operational test.

## 4. Score Reporting

### 4.1 Scoring and Weighting

Of the 88 items in an administration of the Versant Pro – Speaking test, 83 responses are used in the automatic scoring. The first item response of each task type in the test is considered a practice item and is not incorporated into the final score (except for the Read Aloud, Story Retelling, and Passage Comprehension sections).

The Versant Pro – Speaking score report is comprised of an Overall score and five diagnostic subscores (Sentence Mastery, Vocabulary, Fluency<sup>2</sup>, Pronunciation, and Listening Comprehension).

**Overall:** The Overall Score of the test represents the ability to understand spoken English and speak it intelligibly at a native-like conversational pace on everyday and workplace topics. Scores are based on a weighted combination of five sub-scores. Scores are reported in the range from 20 to 80.

**Sentence Mastery:** Sentence Mastery reflects how well the candidate understands and produces a variety of sentence structures in spoken English. The score is based on the ability to

---

<sup>2</sup> Within the context of language acquisition, the term “fluency” is sometimes used in the broader sense of general language mastery. In the narrower sense used in the Versant Pro - Speaking score reporting, “fluency” is taken as a component of oral proficiency that describes certain characteristics of the observable performance. Following this usage, Lennon (1990) identified fluency as “an impression on the listener’s part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently” (p. 391). In Lennon’s view, surface fluency is an indication of a fluent process of encoding. The Versant Pro - Speaking fluency subscore is based on measurements of surface features such as the response latency, speaking rate, and continuity in speech flow, but as a constituent of the Overall score it is also an indication of the ease of the underlying encoding process.

use accurate and appropriate words and phrases in meaningful sentences.

**Vocabulary:** Vocabulary reflects how well the candidate understands and produces a wide range of words in spoken English from everyday and workplace situations. The score is based on the familiarity with the meanings of common words and their use in connected speech.

**Fluency:** Fluency reflects how well the candidate uses appropriate rhythm, phrasing, and timing when speaking English. The score is based on the ability to speak smoothly and naturally at a conversational pace.

**Pronunciation:** Pronunciation reflects how well the candidate produces English consonants, vowels, words and phrases in an intelligible, native-like manner. The score is based on the ability to correctly articulate the sounds of English in connected speech.

**Listening Comprehension:** Listening reflects how well the candidate understands specific details and main ideas from everyday and workplace 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.

Figure 2 illustrates which sections of the test contribute to each of the five subscores. Each vertical rectangle represents a response from a candidate. The items that are not included in the automatic scoring are shown in blue. These include the first item in Repeats, Short Answer Questions, Sentence Builds, Response Selection, and Conversations.

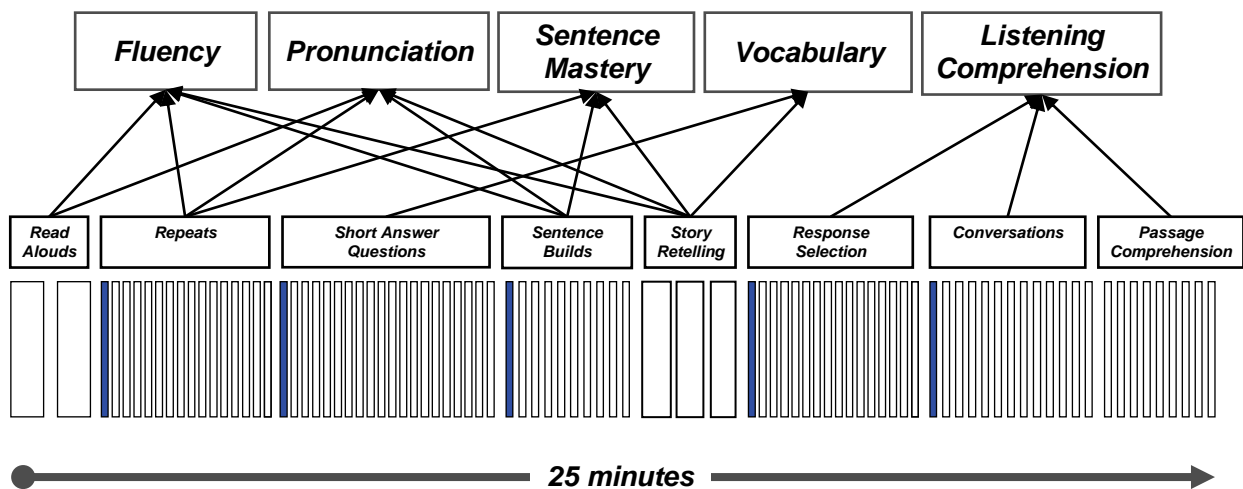


Figure 2. Relation of subscores to item types.

A multi-method, multi-trait approach is taken to ensure that each subscore is reliable and generalizable. The fluency subscore, for example, is derived from performance on four different tasks.

The Versant Pro – Speaking Overall score is a weighted combination of all the subscores. Table 2 shows how the five subscores are weighted to achieve an Overall score.

Table 2. Subscore Weighting in Relation to Versant Pro – Speaking Overall Score.

Subscore	Weighting
Sentence Mastery	20 %
Vocabulary	20 %
Fluency	20 %
Pronunciation	20 %
Listening Comprehension	20 %
<b>Overall Score</b>	<b>100 %</b>

The subscores are based on three different aspects of language performance: a knowledge aspect (the content of a response), a control aspect (the manner in which a response is said), and a comprehension aspect (the extent to which a response reflects the understanding of a listening stimulus). The five subscores reflect these aspects of language performance where Sentence Mastery and Vocabulary are associated with content, Fluency and Pronunciation are associated with manner of speaking, and Listening Comprehension is associated with comprehension. The content accuracy dimension counts for 40% of the Overall score and indicates whether or not the candidate understood the prompt and responded in grammatically accurate sentences and/or with appropriate content. The manner-of-speaking scores count for an additional 40% of the Overall score, and indicate whether or not the candidate speaks in a native-like manner. The remaining 20% of the Overall score comes from the comprehension score and indicates whether or not the candidate understood the spoken material. In smooth, successful communication, it is essential to be able to understand what is being said or asked in a stream of speech. Furthermore, producing accurate lexical and structural content is important, but excessive attention to accuracy can lead to disfluent speech production and can also hinder oral communication; on the other hand, inappropriate word usage and misapplied syntactic structures can also hinder communication. Because successful communication depends on these three dimensions, Versant Pro – Speaking is designed to assess each of them.

The Ordinate automated scoring system scores both the content (including the content of the responses to the listening items) and manner-of-speaking subscores using a speech recognition system that is optimized based on non-native English spoken response data collected during the field test. The content subscores are derived from the correctness of the candidate’s response and the presence or absence of expected words in correct sequences. The manner-of-speaking subscores (Fluency and Pronunciation, as the control dimension) are 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. In order to produce valid scores, during the test development stage, these measures were automatically generated on a sample set of utterances (from both native and non-native speakers) and were then scaled to match human ratings.

## 4.2 Score Use

Once a candidate has completed a test, the Versant testing system analyzes the spoken performances and posts the scores at [www.VersantTest.com](http://www.VersantTest.com). Test administrators and score users can then view and print out the test results from a password-protected section of the website called ScoreKeeper.

Score users of Versant Pro – Speaking may be business organizations, educational and government institutions. Business organizations may use Versant Pro – Speaking scores as part of the screening,

hiring, selection, or promotion process. Within a pedagogical research setting, Versant Pro – Speaking scores may be used to evaluate the level of spoken English skills of individuals entering into, progressing through, and leaving English language courses.

The Versant Pro – Speaking score scale covers a wide range of abilities in spoken English communication. In most cases, score users must decide what Versant Pro – Speaking score is considered a minimum requirement in their context (i.e., a cut score). Score users may wish to base their selection of an appropriate cut score on their own localized research. Pearson can provide a Benchmarking Kit and assistance in establishing cut scores.

## 5. Field Test

### 5.1 Data Collection

Both native speakers and non-native speakers of English were recruited as participants from August 2009 through November 2009 to take a prototype data collection version of the Versant Pro – Speaking test. The purposes of this field testing were 1) to validate operation of the test items with both native and non-native speakers; 2) to calibrate the difficulty of each item based on a large sample of candidates at various levels and from various first language backgrounds; 3) to collect sufficient English speech samples to train and optimize the automatic speech processing system; and 4) to develop automatic scoring models for spoken English.

#### 5.1.1 Native Speakers

A total of 73 educated adult native English speakers were recruited. Most were from the U.S. with a few from the U.K. and Australia. Most of them took the test multiple times on the CDT platform producing a total of 706 completed tests. Each test was comprised of a unique set of items, so items did not overlap between the tests. The mean age of the native speaker sample was 35.6 and the male:female ratio was 31:69.

While Versant Pro – Speaking is specifically designed for non-native candidates, 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.

#### 5.1.2 Non-Native Speakers

A total of 973 non-native candidates were recruited from various countries representing both university students and working professionals. All the data collection tests were taken on the CDT platform. Most candidates took both Versant Pro - Speaking and Versant Pro – Writing. Almost all the candidates took the test only once.

Based on the country of origin, there were a total of 46 countries represented in the field test, but the majority of the data were collected in Argentina, China, Germany, India, Italy, Japan, Korean, the Philippines, Spain, and Taiwan. A total of 46 different languages were reported. The male:female ratio was 50:47 with 3% of the candidates being unknown. The mean age was 28.9.

Versant Pro – Speaking shares four tasks with the Versant English Test: Repeats, Short Answer Questions, Sentence Builds, and Story Retelling. Therefore, many items in these tasks were seeded into the Versant English Test item bank in order to collect additional responses, facilitate the data collection process, and to calibrate the two tests to a common scale.

Table 3 summarizes the demographic information of both the native and non-native candidates who took the data collection version of Versant Pro – Speaking.

Table 3. Description of Participants in the Field Testing Whose Responses were Used to Develop Automated Scoring Models (n=1,046).

	Native	Non-Native
<b>Number of Participants</b>	73	973
<b>Male : Female ratio</b>	31% : 63% Unknown = 6%	50% : 47% Unknown = 3%
<b>Age Range</b>	20 - 73 mean = 35.6	19 – 67 mean = 28.9
<b>Languages</b>	English (U.S., U.K., and Australia)	Angami, Arabic, Assamese, Bengali, Cantonese, Catalan, Cebuano, Chavacano, Chinese, Czech, Dutch, Farsi, Filipino, Fookien, French, German, Gujarati, Haryanvi, Hindi, Indonesian, Japanese, Kalenjin, Kannada, Korean, Maithali, Malayalam, Manipuri, Marathi, Marwadi, Oriya, Portuguese, Punjabi, Rongmei, Russian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Urdu, Vietnamese, Visayan, Waray-waray, Yoruba

## 6. Data Resources for Scoring Development

### 6.1 Data Preparation

During the field test of Versant Pro – Speaking, more than 150,000 responses were collected from natives and English learners; of those, 19,671 responses were multiple-choice non-speech responses from Response Selection. Subsets of the response data were transcribed and were presented to trained raters for developing the automatic scoring models.

#### 6.1.1 Transcription

Both native and non-native responses were transcribed by native speakers of English in order to train an automatic speech recognition system optimized for non-native speech patterns. The majority of the transcribers had a degree in a language-related field such as linguistics, language studies, or English. Transcribers underwent rigorous training prior to the task and the quality of their transcriptions was regularly reviewed for accuracy. A total of 30,718 transcriptions were produced for native responses and 87,746 transcriptions were produced for non-native responses.

#### 6.1.2 Expert Human Rating

Selected item responses to Story Retellings from a subset of candidates were presented to nine educated native English speakers to be judged for content accuracy and vocabulary usage to make the

Story Retelling task automatically scoreable. Before the native speakers began rating responses, they received training in how to evaluate responses according to analytical and holistic rating criteria. All raters held a master's degree in either linguistics or TESOL.

The raters logged in to a web-based rating system and evaluated transcriptions of Story Retelling responses, one at a time, for content and vocabulary. The raters' judgments were based on transcriptions instead of recorded spoken responses in order to minimize confounding effects - that is, to ensure that pronunciation or fluency qualities would not affect the evaluation of content and vocabulary. Rating stopped when each item had been judged by at least two raters. For pronunciation and fluency models, the models developed for the Versant English Test were used because those pronunciation and fluency models were trained on a much larger sample of non-native candidates and have proven to be content independent and very robust.

## 7. Validation

### 7.1 Validity Study Design

A series of validity analyses were conducted to examine five aspects of the Versant Pro – Speaking test scores:

#### Internal Validity

1. **Reliability:** whether or not the Versant Pro – Speaking test is reliable and internally consistent,
2. **Dimensionality:** whether or not the five different subscores of the Versant Pro – Speaking test are sufficiently distinct,
3. **Accuracy:** whether or not the automatically scored Versant Pro – Speaking test scores are comparable to the scores that human listeners and raters would assign,
4. **Differentiation among known populations:** whether or not Versant Pro – Speaking scores reflect expected differences and similarities among known populations (e.g., natives vs. English learners),

#### Concurrent Validity

5. **Relation to scores of tests with related constructs:** how closely do Versant Pro – Speaking scores predict the reliable information in scores of a well-established English test for a workplace context (e.g., TOEIC).

#### 7.1.1 Validation Sample

A total of 124 participants were recruited for a series of validation analyses. These validation participants were recruited separately from the field test candidates. Care was taken to ensure that the training dataset and validation dataset did not overlap for independent validation analyses. This means that the spoken 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 Versant Pro – Speaking. A total of five native speakers were included in the validation dataset. Table 4 below summarizes the demographic information of the validation participants.

Table 4. Description of Participants Used to Validate the Scoring Models and Estimate Test Reliability (n=124).

<b>Number of Participants</b>	124 (including 5 native speakers)
<b>Male : Female ratio</b>	44% : 56%
<b>Age Range</b>	19 - 66 mean = 30.4
<b>Languages</b>	Arabic, Chinese, English, Filipino, French, German, Hindi, Italian, Japanese, Korean, Malayalam, Russian, Spanish, Tagalog, Tamil, Telugu, Visayan

## 7.2 Internal Validity

### 7.2.1 Descriptive Statistics

The mean Overall score of the validation sample was 49.57 with a standard deviation of 15.15 (on a scale of 20-80). Table 5 summarizes some descriptive statistics for the validation sample.

Table 5. Descriptive Statistics for the Validation Dataset (n=124).

<b>Measure</b>	<b>Statistic</b>
<b>Mean</b>	49.57
<b>Standard Error</b>	1.36
<b>Median</b>	48.07
<b>Standard Deviation</b>	15.15
<b>Sample Variance</b>	229.54
<b>Kurtosis</b>	-0.63
<b>Skewness</b>	0.33

### 7.2.2 Test Reliability

To understand the consistency and accuracy of Versant Pro – Speaking Overall scores and the distinctness of the subscores, the following were examined: the standard error of measurement of the Versant Pro – Speaking Overall score; the reliability of Versant Pro – Speaking (split-half reliability); the correlations between the Versant Pro – Speaking Overall score and its subscores, and between pairs of subscores; comparison of machine-generated Versant Pro – Speaking scores with listener-judged scores of the same Versant Pro – Speaking tests. These qualities of consistency and accuracy of the test scores are the foundation of any valid test.

The Standard Error of Measurement (SEM) provides an estimate of the amount of error, due to unreliability, in an individual’s observed test score and “shows how far it is worth taking the reported score at face value” (Luoma, 2003: 183). The SEM of the Versant Pro – Speaking Overall score is 2.3.

Score reliabilities were estimated by the split-half method. Split-half reliability was calculated for the Overall score and all subscores. The split-half reliabilities use the Spearman-Brown Prophecy Formula to correct for underestimation. The split-half reliabilities were calculated for both the listener-judged scores and the machine-generated scores. The reliability coefficients are summarized in Table 6. The human scores were calculated from human transcriptions (for the Sentence Mastery, Vocabulary and Listening Comprehension subscores) and human judgments (for the Pronunciation and Fluency subscores). Table 6 compares the same individual performances, scored by careful human rating in one case and by independent automatic machine scoring in the other case. The values in Table 6 suggest that there is sufficient information in a Versant Pro - Speaking item response set to extract reliable information, and that the effect on reliability of using the Ordinate speech recognition technology, as opposed to careful human rating, is quite small. The high reliability score is a good indication that the computerized assessment will be consistent for the same candidate assuming there are no changes in the candidate’s language proficiency level.

Table 6. Split-half Reliabilities of Versant Pro - Speaking Human Scoring versus Machine Scoring (n=124).

Score	Split-half Reliability for Human Scores	Split-half Reliability for Machine Scores
<b>Overall</b>	0.99	0.98
<b>Sentence Mastery</b>	0.96	0.92
<b>Vocabulary</b>	0.94	0.88
<b>Fluency</b>	0.99	0.96
<b>Pronunciation</b>	0.99	0.97
<b>Listening Comprehension</b>	0.93	0.90

The reliability for the Vocabulary subscore is lower than the reliability of the other subscores. This may be because the variability in candidates’ vocabulary knowledge across test items is large in relation to the vocabulary presented in the Short Answer Question items and Story Retelling items that are used as the basis for the Vocabulary subscore.

### 7.2.3 Dimensionality: Correlations Among Subscores

Each subscore 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 subscores 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 subscores were to correlate perfectly with one another, then the subscores might not be measuring different aspects of facility with the language.

Table 7 presents the correlations among the Versant Pro – Speaking subscores and the Overall score for the same validation sample of 124 candidates, which includes five native English speakers.

Table 7. Inter-correlation between Subscores on the Versant Pro – Speaking Test (n=124).

	Sentence Mastery	Vocab.	Fluency	Pronunciation	Listening Comp.	Overall
Sentence Mastery	-	0.86	0.80	0.74	0.77	<b>0.91</b>
Vocabulary		-	0.80	0.82	0.82	<b>0.92</b>
Fluency			-	0.80	0.80	<b>0.94</b>
Pronunciation				-	0.71	<b>0.88</b>
Listening Comprehension					-	<b>0.91</b>

As expected, test 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. However, the correlations between the subscores are significantly below unity, which indicates that the different scores measure different aspects of the test construct.

Figure 3 illustrates the relationship between two relatively independent machine scores (Sentence Mastery and Fluency) for the validation sample (n=124). These machine scores are calculated from a subset of responses that are mostly overlapping (Repeats, Sentence Builds, and Story Retellings for Sentence Mastery and Read Alouds, Repeats, Sentence Builds, and Story Retellings for Fluency). Although these measures are derived from overlapping sets of responses, the subscores clearly extract distinct measures from these responses. For example, candidates with Fluency scores in the 30-50 range have Sentence Mastery scores that are spread roughly evenly over the whole 20-80 score range.

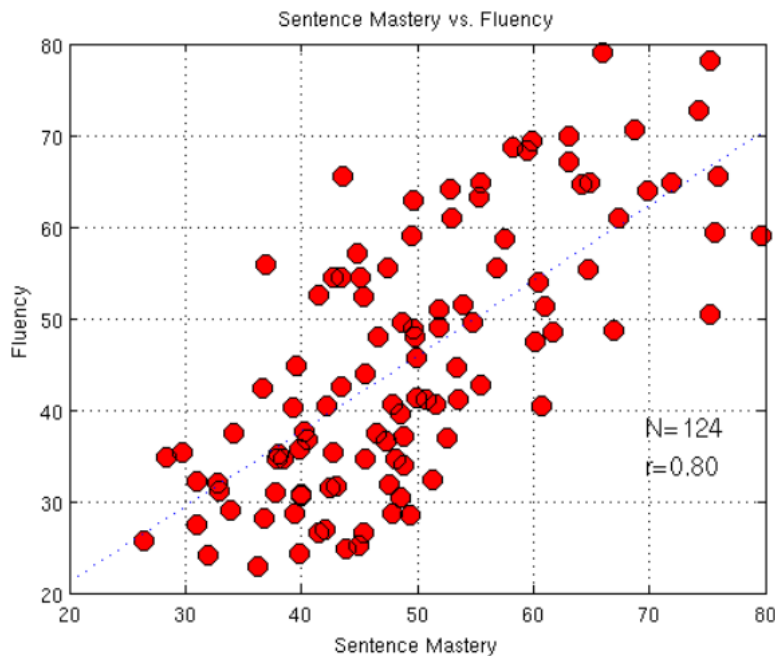


Figure 3. Sentence Mastery vs. Fluency scores for the validation sample (r = 0.80)

### 7.2.4 Machine Accuracy

Another analysis for internal quality involved comparing scores from the Versant Pro – Speaking test, which uses Ordinate’s speech processing technologies, versus careful human transcriptions and human judgments from expert raters.

Table 8 presents Pearson Product-Moment correlations between machine scores and human scores, when both methods are applied to the same performances on the same Versant Pro – Speaking responses. The candidate sample is the same set of 124 validation candidates that was used in the reliability and subscore analyses. Correlations presented in Table 8 suggest that scoring a Versant Pro – Speaking test by machine will yield scores that generally correspond as they should with human ratings. Among the subscores, the human-machine relation is closer for the content accuracy subscores (Sentence Mastery and Vocabulary) than for the manner-of-speaking subscores (Fluency and Pronunciation), but the relation is close for all five subscores. At the Overall score level, Versant Pro – Speaking machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and multiple independent human judgments.

Table 8. Correlations between Human and Machine Scoring of Versant Pro – Speaking Responses (n = 124).

Score Type	Correlation
<b>Overall</b>	0.95
<b>Sentence Mastery</b>	0.93
<b>Vocabulary</b>	0.95
<b>Fluency</b>	0.85
<b>Pronunciation</b>	0.84
<b>Listening Comprehension</b>	0.96

### 7.2.5 Differentiation among Known Populations

The next validity analysis examined whether or not the Versant Pro – Speaking scores reflect expected differences between native English speakers and English language learners. This means that Overall scores from learners should distribute over the score range according to their spoken English ability, whereas the native speakers should receive high Overall scores.

Overall scores from 28 native speakers and 987 non-native speakers representing a range of native languages were compared. Figure 4 presents the score distributions of Overall scores for the native and non-native speakers in the form of histograms. The results show that native speakers of English consistently obtain high scores on Versant Pro – Speaking (in red). All native test-takers scores fall into the last score bin of 76-80. On the other hand, learners of English as a second or foreign language are normally distributed over a wide range of scores. The Overall scores show effective separation between native and non-native candidates.

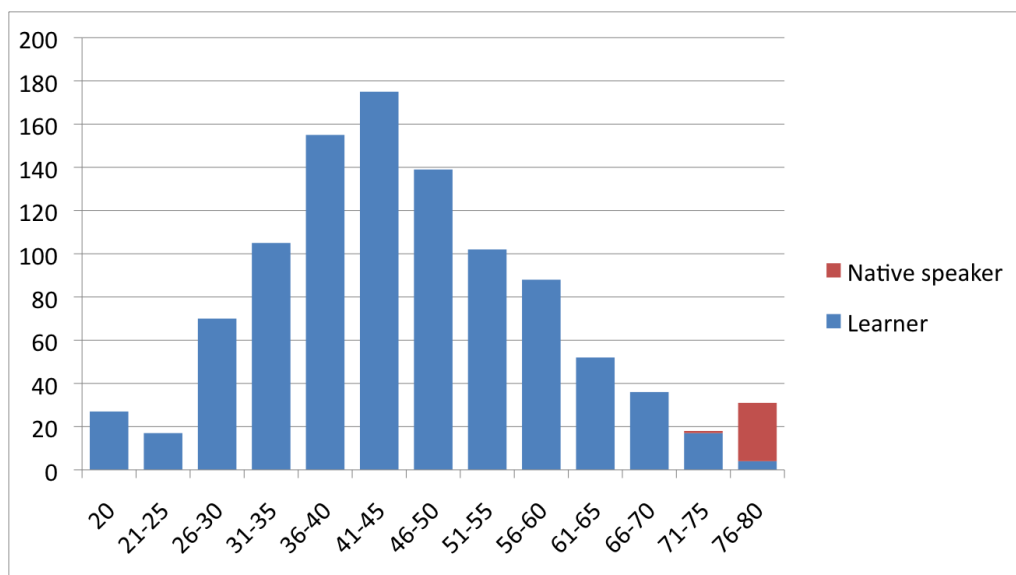


Figure 4. Histograms of Versant Pro - Speaking Overall scores for the native and non-native groups (native n=28 and non-native n=987).

## 7.3 Concurrent Validity

### 7.3.1 Versant Pro – Speaking and TOEIC

One important goal of the validity studies is to understand how Versant Pro – Speaking relates to other measures of English proficiency. Since the Versant Pro – Speaking test has an emphasis on workplace English, it would be most sensible to explore a relationship with another well-known workplace English test. For this reason, a study was undertaken to compare the automatically derived Versant Pro – Speaking Overall scores with the Test of English for International Communication (TOEIC).

TOEIC is comprised of a Listening and Reading test and a Speaking and Writing test. In this concurrent validation study, the data with the TOEIC Listening and Reading test was collected because it is the more widely-used test. The TOEIC Listening Reading test is claimed to measure “a non-native speaker’s listening and reading skills in English as these skills are used in the workplace. The test was developed about 30 years ago as a measure of receptive language skill and has been widely accepted and used worldwide.” (Liao, Qu, & Morgan, 2010). The listening and reading subscores are both reported in the range of 5 to 495 for a total score between 10 and 990.

#### Method

The study was conducted between November 2009 and February 2010. The participants were 28 Japanese and 27 South Koreans who represented a mix of full-time students and working professionals. Of the 55 participants, 26 were male and 29 female with a mean age of 24. The participants were recruited by agents in Japan and Korea acting on Pearson’s behalf (a university professor and two business professionals).

The participants took both the Versant Pro – Speaking and TOEIC tests with a gap between sittings of no less than 30 days. All participants were asked to take a shorter version of Versant Pro – Speaking as

a demo test so their resulting performance would more closely relate to their proficiency levels, rather than reflect their unfamiliarity with the Versant Pro – Speaking test. All participants were already familiar with the format of the TOEIC test. They took Versant Pro – Speaking individually at their home, school, or workplace. The TOEIC tests were administered during the official test administrations. No institutional TOEIC tests were used.

## Results

The inter-correlation matrix between the subscores of each test is given in Table 9. The values across all subscores are at or above  $r=0.67$ . Not surprisingly, the highest correlation coefficients (0.96 and 0.91) exist between subscores (or modules) and the overall scores for the same test. This is true for both Versant Pro and TOEIC. When the overall scores from Versant Pro – Writing and Versant – Pro Speaking are combined, the correlation between the Versant Pro total and the TOEIC total is  $r=0.78$ .

Table 9. Pearson Correlation Coefficients for Versant Pro and TOEIC (n=55).

	TOEIC Reading	TOEIC Listening	TOEIC Total	Versant Pro – Speaking
TOEIC Reading	-			
TOEIC Listening	0.84	-		
TOEIC Total	0.96	0.96	-	
Versant Pro – Speaking	0.67	0.71	0.72	-
Versant Pro Total <sup>3</sup>	0.75	0.76	0.78	0.91

Though the sample size is small, this matrix shows an expected pattern of relationships among the subscores of the tests, bearing in mind they all relate to English language ability but assess different dimensions of that ability.

Versant Pro – Speaking and TOEIC Total correlated moderately at  $r=0.72$ , as shown in Figure 5, indicating that there is general English ability as a covariance, but that these tests measure different aspects of language performance. Versant Pro – Speaking correlated higher with TOEIC Listening ( $r=0.71$ ) than with TOEIC Reading ( $r=0.67$ ), which is expected because Versant Pro – Speaking is designed to measure listening and speaking abilities rather than reading ability.

<sup>3</sup> The Versant Pro Total score represents the combination of the overall scores from Versant Pro – Speaking and Versant Pro – Writing.

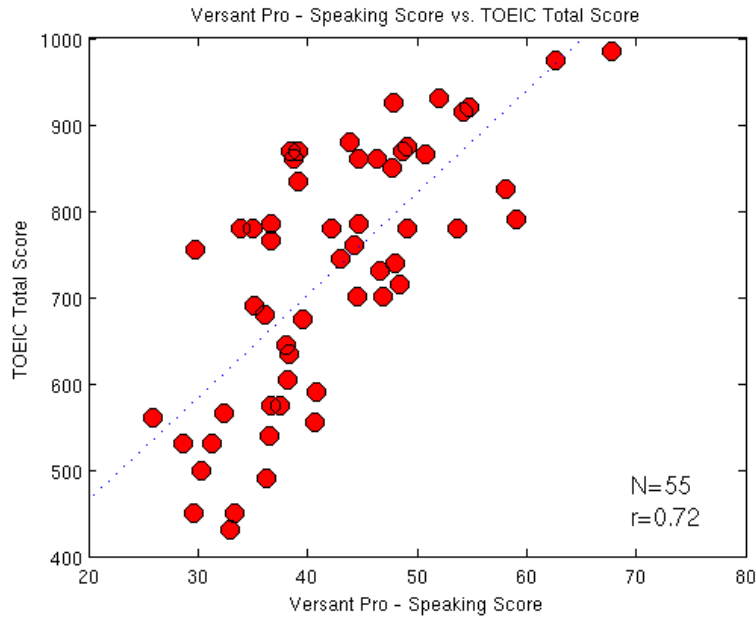


Figure 5. Scatterplot showing the relationship between Versant Pro – Speaking and TOEIC (n=55).

### 7.3.2 Versant Pro – Speaking and Versant English Test

A study was conducted to explore the relation between scores on Versant Pro – Speaking and the Versant English Test (VET). 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. If there is a close relation between Versant Pro – Speaking and the Versant English Test, it then follows that Versant Pro – Speaking also measures what it claims to measure – facility in spoken English.

The analysis involved taking the Versant Pro –Speaking validation set of 124 candidates and extracting their performances on the parts of the test that share similarities with the Versant English Test (i.e. Repeats, Shorts Answer Questions, Sentence Builds, and Story Retells). These responses were processed through the automated scoring algorithms for the Versant English Test and the scores were compared to the Versant Pro – Speaking scores, as shown in Figure 6.

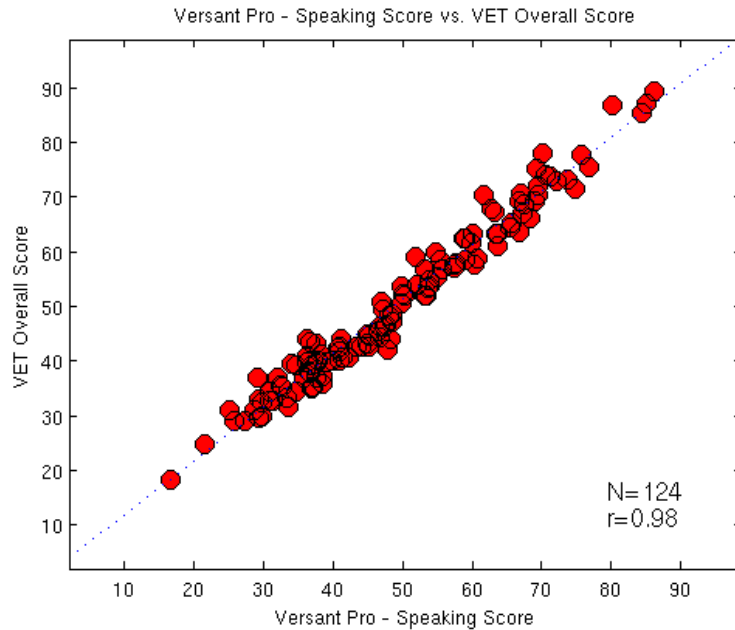


Figure 6. Scatterplot showing the relation between Versant Pro – Speaking scores and Versant English Test scores.

As can be seen, there is a strong relation between the two sets of scores ( $r=0.98$ ). This strong relation between Versant Pro – Speaking and the Versant English Test supports the claim that Versant Pro – Speaking measures the intended construct.

Although the relation between the two tests is strong when investigated at the entire sample level, there are clear individual differences between candidates, as shown in Table 10. Versant Pro – Speaking and Versant English Test scores correspond to one another at essentially a one-to-one correspondence (e.g. 47 on one test is equal to 47 on the other test), but the scores of individuals vary depending on the personal skill set.

Table 10. Score comparison between Versant Pro – Speaking and the Versant English Test.

Proportion of candidates (n=124)	Score point difference between Versant Pro - Speaking and VET
1 %	$\geq 9$ points difference
10 %	$\geq 6$ points difference
31 %	$\geq 4$ points difference
58 %	$\leq 3$ points difference

This difference in scoring between the Versant Pro – Speaking and the Versant English Test is largely due to the impact of the Listening Comprehension subscore in Versant Pro – Speaking. The correlation coefficient of Listening Comprehension with the other four subscores of the test combined (Sentence Mastery, Vocabulary, Fluency and Pronunciation) is  $r=0.84$ . This high correlation reveals that Listening Comprehension shares common variance with the other subscores of the test but also contributes unique variance to the Overall score.

### 7.3.3 Versant Pro – Speaking and CEFR Level Estimates

Because of the high correlation ( $r=0.98$ ) between Versant Pro – Speaking and the Versant English Test, the results from a previous study mapping Versant English scores onto the CEFR levels have been applied to Versant Pro – Speaking. That is, the established Versant English score ranges aligned with the CEFR levels have been used for Versant Pro – Speaking, as shown in Table 11. The method used to create the mappings is described in the *Can-Do Guide*. Please contact Pearson for this report.

Table 11. Mapping of CEFR Levels with Versant Pro – Speaking Scores.

Versant Pro – Speaking Score Range 20 – 80	CEFR A1-C2
20-25	<A1
26-35	A1
36-46	A2
47-57	B1
58-68	B2
69-78	C1
79-80	C2

## 8. Conclusion

This report has provided details of the test development processes and validity evidence for the Versant Pro – Speaking test. The information is available for test users to make an informed interpretive judgment as to whether test 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, namely:

- the items were written to specifications and were subjected 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 scores, item weights and scoring logic are clearly explained; and
- the items were widely field tested on a representative sample of candidates.

This report provides empirical evidence demonstrating that Versant Pro – Speaking scores are structurally reliable indications of a candidate’s ability in spoken English and are suitable for high-stakes decision-making.

## 9. About the Company

**Pearson:** Pearson’s Knowledge Technologies group 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 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 scoring report generators linked to the Internet. Versant Pro – Speaking is the result of years of research in speech recognition, statistical modeling, linguistics, and testing theory. The Versant patented technologies are applied to Pearson’s own language tests such as the Versant English series and also to customized tests. Sample projects include assessment of spoken English, assessment of spoken aviation 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.VersantTest.com](http://www.VersantTest.com)).

## 10. References

- 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.
- 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.
- Carroll, J.B. (1961). *Fundamental considerations in testing for English language proficiency of foreign students*. Testing. Washington, DC: Center for Applied Linguistics.
- Carroll, J.B. (1986). *Second language*. In R.F. Dillon & R.J. Sternberg (Eds.), *Cognition and Instructions*. Orlando FL: Academic Press.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: 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>.

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.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387-412.

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.

Liao, C.-w., Qu, Y., and Morgan, R. (2010). *The Relationships of Test Scores Measured by the TOEIC Listening and Reading Test and TOEIC Speaking and Writing Tests (TC-10-13)*. Princeton, NJ: Educational Testing Service.

Luoma, S. (2003). *Assessing Speaking*. Cambridge, 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, 217-228.

Ordinate (2003). *Ordinate SET-10 Can-Do Guide*. Menlo Park, CA: Author.

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

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

## II. Appendix A: Test Paper




Instructions and general introduction to test procedures.



### Test Instructions

#### Please read this before taking the test

Versant tests are automated spoken language tests that are taken on the telephone or computer. If you would like to listen to a sample test, purchase a practice test, or view the test score after taking the test (if applicable), please visit [www.VersantTest.com](http://www.VersantTest.com)

Part	Instructions
<b>Before the Test</b>	<ul style="list-style-type: none"> <li>Carefully read this instruction page and the test paper. You may use a dictionary or ask someone for help if there are words or sentences that you don't understand.</li> <li>Choose a quiet location with a landline phone where you will not be interrupted during the test.</li> <li>Do not use a cordless phone, cellular phone, or VoIP phone (e.g., Skype™ or PC-to-phone services). Newer phones are generally better than older phones. Make sure that the phone is set to tone and not pulse.</li> </ul>
<b>Beginning the Test</b>	<ul style="list-style-type: none"> <li>To begin the test, call the phone number on the test paper using a landline push-button telephone.</li> <li>A recorded examiner's voice will guide you through each section of the test.</li> <li>Enter your Test Identification Number using the telephone keypad when the examiner's voice asks you to do so. This number is printed on the top right of your test paper.</li> <li>The examiner's voice will then ask you two questions: your name, and the city and the country you are calling from. If you are speaking too loudly or too quietly, the examiner's voice will tell you.</li> <li>The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.</li> </ul>
<b>During the Test</b>	<ul style="list-style-type: none"> <li>Hold the phone close to your mouth as shown in the picture below.</li> </ul> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>NO Too low, too far away</p> </div> <div style="text-align: center;">  <p>YES In front of mouth</p> </div> <div style="text-align: center;">  <p>YES A good distance</p> </div> </div> <ul style="list-style-type: none"> <li>Answer all questions smoothly and naturally in a clear, steady voice.</li> <li>If you don't know the proper way to respond to a test item, you can remain silent or say, "I don't know."</li> <li>Do not take notes or write during the test.</li> <li>When you hear, "Thank you for completing the test", you may hang up.</li> <li>If you wish, you may answer the optional questions at the end of the test. Your personal information will be kept anonymous.</li> </ul>

PEARSON

© 2011 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

**Side I of the Test Paper:** Individualized test form (unique for each candidate) showing Test Identification Number, Part A: passages to read, and examples for Parts B, C, and D.



## Versant Pro - Speaking

REMINDER: The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.

**Call: 1-415-738-3800**

Test Identification Number (TIN)

**1234 5678**

Expires: May 20, 2011

Thank you for calling the Versant testing system.  
 Please enter your Test Identification Number on the telephone keypad.  
 Now, please say your name.  
 Now, please say the city and country you are calling from.  
 Now, please follow the instructions for Parts A through H.

PART	TASK	TEST DETAILS
<b>A</b>	<b>Read Aloud</b>	<p>Read the passage aloud smoothly and naturally in a clear voice. You will be stopped after 30 seconds. This is not a speed reading test. You may not be able to finish reading the entire passage, but that is okay. When your time is up, you will automatically move on to the next item.</p> <p>1. We have several offices for rent in a large office building. The building is surrounded by trees. All offices have private balconies and hardwood floors. There are many features including an outdoor eating area and a shower. The location is within a few steps of many shops and cafes.</p>
<b>B</b>	<b>Repeat</b>	<p>Please repeat each sentence that you hear.</p> <p>Example: You hear: "Prices are going up."                      You say: "Prices are going up."</p>
<b>C</b>	<b>Questions</b>	<p>Please give a simple answer to the questions.</p> <p>Example 1: You hear: "Which carries more people, a motorcycle or an airplane?"                      You say: "airplane" or "an airplane"</p> <p>Example 2: You hear: "What month comes after January?"                      You say: "February"</p>
<b>D</b>	<b>Sentence Builds</b>	<p>Please rearrange the word groups into a sentence.</p> <p>Example: You hear: "finished by April"..."must be"..."our projects"                      You say: "Our projects must be finished by April."</p>

Next Page

**PEARSON**

01 - 11111 - 1

© 2010 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

**Side 2 of the Test Paper:** Individualized test form (unique for each candidate) showing examples for Parts E, F, G, and H.



PART	TASK	TEST DETAILS
<b>E</b>	<b>Story Retellings</b>	<i>You will hear a short story. After the story, you will have 30 seconds to retell the story in English as best you can. Try to retell as much of the story as you can, including the situation, characters, actions, and ending.</i>
<b>F</b>	<b>Response Selection</b>	<p><i>You will hear a sentence followed by three possible responses. Say the letter A, B, or C to answer. You will have 5 seconds to answer. Say clearly "A" or "B" or "C" and only once.</i></p> <p>Example: You hear the sentence:            "What time is it now?"            Then, you hear three possible responses:            "A. I like reading newspapers."            "B. Food is getting expensive."            "C. It's nine o'clock."            Then, you say "C".  <i>Remember, just say A, B, or C clearly and only once.</i></p>
<b>G</b>	<b>Conversations</b>	<p><i>You will hear a conversation between two people, followed by a question. Give a short, simple answer to the question.</i></p> <p>Example: You hear: Speaker 1: "Lucy, can you come to the office early tomorrow?"            Speaker 2: "Sure, what time?"            Speaker 1: "7:30 would be great."            Question: "What will Lucy have to do tomorrow morning?"            You say: "Go to the office early" or "She will go to the office at 7:30."</p>
<b>H</b>	<b>Passage Comprehension</b>	<i>You will hear a story, followed by three questions. Please give a short, simple answer to each of the questions. Your answer could be a few words or a very short sentence.</i>

*Thank you for completing the test.*

**PEARSON**

01 - 11111 - 2

© 2010 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).



## About Us

The Knowledge Technologies group of 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.

The Knowledge Technologies group is part of Pearson, the international media company, whose businesses also include the Financial Times Group and the Penguin Group.

### Pearson

299 S. California Avenue  
Suite 300  
Palo Alto, California 94306  
USA

4940 Pearl East Circle  
Suite 200  
Boulder Colorado 80301  
USA



**Contact Us**  
To try a sample test or get  
more information, contact us at:

US: 800.211.8378  
Int'l: +1 650.470.3505  
sales@pearsonkt.com

Or visit us online at:  
[www.VersantTest.com](http://www.VersantTest.com)

Pearson now includes Ordinate products and services.

© 2011 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.

Version 1211B